



UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Biologia

ANDRÉ RICARDO OLIVEIRA CONSON

MAPEAMENTO GENÉTICO-MOLECULAR E ANÁLISE DE QTLs
ASSOCIADOS AO CRESCIMENTO EM *HEVEA BRASILIENSIS*
UTILIZANDO MARCADORES MICROSSATÉLITES E SNPs

GENETIC MAPPING AND QTL ANALYSIS OF GROWTH-RELATED
TRAITS IN *HEVEA BRASILIENSIS* USING MICROSATELLITE AND
SNP MARKERS

Campinas

2016

ANDRÉ RICARDO OLIVEIRA CONSON

MAPEAMENTO GENÉTICO-MOLECULAR E ANÁLISE DE QTLs ASSOCIADOS AO
CRESCIMENTO EM *HEVEA BRASILIENSIS* UTILIZANDO MARCADORES
MICROSSATÉLITES E SNPs

GENETIC MAPPING AND QTL ANALYSIS OF GROWTH-RELATED TRAITS IN
HEVEA BRASILIENSIS USING MICROSATELLITE AND SNP MARKERS

Tese apresentada ao Instituto de Biologia da
Universidade Estadual de Campinas como parte
dos requisitos exigidos para a obtenção do título
de Doutor em Genética e Biologia Molecular na
área de Genética Vegetal e Melhoramento

*Thesis presented to the Institute of Biology of the
University of Campinas in partial fulfillment of
the requirements for the degree of Doctor in
Genetics and Genetic Breeding*

ESTE ARQUIVO DIGITAL CORRESPONDE À
VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO
ANDRÉ RICARDO OLIVEIRA CONSON, E
ORIENTADA PELA PROF(A). DR(A). ANETE
PEREIRA DE SOUZA

Orientadora/Supervisor: Prof.^a Dr.^a Anete Pereira de Souza

Coorientadora/Co-supervisor: Dr.^a Livia Moura de Souza

Campinas

2016

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2007/50562-4 e 2012/50491-8;
CNPq, 562979/2010-7, 478701/2012-8 e 162629/2015-0

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Gustavo Lebre de Marco - CRB 8/7977

Conson, André Ricardo Oliveira, 1987-
C765m Mapeamento genético-molecular e análise de QTLs associados ao
crescimento em *Hevea brasiliensis* / André Ricardo Oliveira Conson. –
Campinas, SP : [s.n.], 2016.

Orientador: Anete Pereira de Souza.

Coorientador: Livia Moura de Souza.

Tese (doutorado) – Universidade Estadual de Campinas, Instituto de
Biologia.

1. Seringueira. 2. *Hevea*. 3. Mapeamento cromossômico. 4. Mapeamento
de QTL. 5. Marcadores moleculares. I. Souza, Anete Pereira de. II. Souza,
Livia Moura de. III. Universidade Estadual de Campinas. Instituto de Biologia.
IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Genetic mapping and QTL analysis of growth-related traits in
Hevea brasiliensis

Palavras-chave em inglês:

Rubber plants

Hevea

Chromosome mapping

QTL analysis

Molecular markers

Área de concentração: Genética Vegetal e Melhoramento

Titulação: Doutor em Genética e Biologia Molecular

Banca examinadora:

Anete Pereira de Souza [Orientador]

Gabriel Rodrigues Alves Margarido

Mirian Perez Maluf

Rafael Vasconcelos Ribeiro

Marcelo Mendes Brandão

Data de defesa: 05-12-2016

Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, 05 de dezembro de 2016

COMISSÃO EXAMINADORA

Prof.^a Dr.^a Anete Pereira de Souza

Prof. Dr. Gabriel Rodrigues Alves Margarido

Dr.^a Mirian Perez Maluf

Prof. Dr. Rafael Vasconcelos Ribeiro

Prof. Dr. Marcelo Mendes Brandão

Os membros da comissão examinadora acima assinaram a Ata da Defesa, que se encontra no processo de vida acadêmica do aluno.

“Education is our passport to the future,
for tomorrow belongs to the the people
who prepare for it today”

Malcolm X

Dedicatória

À Luciana, que me ajudou a enfrentar todos os desafios

Agradecimentos

À Prof.^a Dr.^a Anete, por todos os ensinamentos. Também pela amizade, apoio e por ter me recebido com carinho e confiado no meu trabalho.

À minha coorientadora, Dr.^a Livia, por ajudar a me integrar no grupo de seringueira e sempre exigir o melhor de cada um.

À minha esposa Luciana, que durante todos esses anos juntos me ajudou com escolhas de vida, me compreendeu em momentos de dificuldade, e me mostrou como devemos buscar a felicidade.

À minha mãe Rose, que desde quando decidi seguir em frente e buscar meus objetivos profissionais me apoiou. Àquela que me incentivou a buscar o conhecimento.

Ao meu pai Luiz e sua incansável vontade de aprender algo novo. De onde eu tirei toda a animação pela ciência.

Aos meus avós, com quem vivenciei “tempo de qualidade” durante toda a minha vida.

À Fátima, que com carinho tem cuidado dos meus avós.

À família Gimenes, pelos momentos de apoio e por serem minha segunda casa.

À toda a minha família. Pelos momentos prazerosos juntos no fim de ano, ou mesmo em curtos fins de semana. Por saber que todos sempre torceram pelo meu sucesso.

Aos amigos do grupo de seringueira, Livia, Camila, Carla, Guilherme, Luciano, Isabela. Com amigos como esses, fica difícil apenas pensar nos problemas e dificuldades do genoma da seringueira.

Ao Luciano pela parceria nos almoços no RS.

Ao professor Dr. Antônio Augusto Franco Garcia, do Departamento de Genética da ESALQ-USP e seus alunos, Cristiane, Rodrigo e João, pela colaboração e ajuda nas análises estatísticas.

Ao Dr. Vincent Le Guen, do Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), que durante todo o trabalho esteve presente em diversos momentos cruciais, e que me acolheu durante o período de sanduíche em Montpellier.

Ao Prof. Dr. Gabriel Rodrigues Alves Margarido, Prof. Dr. Antônio Augusto Franco Garcia e Dr. Marcelo Molinari pelo desenvolvimento da nova versão do pacote OneMap.

Ao Dr. Paulo de Souza Gonçalves e Dr. Erivaldo José Scaloppi Junior pelo apoio para a obtenção da população de mapeamento e coleta dos dados fenotípicos.

A todos aqueles que passaram pelo laboratório e que compartilharam um pouco da sua experiência comigo. Em especial ao Juvérlande, Carlos, Aline, Patrícia e Danilo.

Aos amigos Maiara e Mário, Murilo e Camila, pelos momentos de alegria e descontração, seja em Campinas, Londrina ou Curitiba.

Ao já companheiro de jornada acadêmica, Caconde, que desde 2006 vem me acompanhando nesse processo de formação.

Aos outros companheiros espalhados por aí desde a formatura, Bruno, Rafaela, André e Vitor por me apoiarem mesmo de longe.

Aos membros da pré-banca e banca, por terem aceito o convite para melhoria deste trabalho.

Às secretárias do Centro de Biologia Molecular, Tânia e Sandra.

À Lourdes e todo o pessoal do Programa de Pós-Graduação em Genética e Biologia Molecular.

À UNICAMP, sempre referência de qualidade na minha vida.

A CAPES e CNPq pelas bolsas de doutorado.

A cooperação científica CAPES/Fundação Agrópolis pela bolsa de doutorado sanduíche.

Resumo

O aspecto emblemático que a seringueira sugere quando se trata de produção de borracha natural no Brasil, é limitado atualmente devido, principalmente, a ausência de clones resistentes a doença chamada mal-das-folhas, causada pelo fungo *Microcyclus ulei*. Clones adaptados a “áreas de escape”, onde níveis epidêmicos da doença não são observados, podem ser uma alternativa para contornar problemas relacionados às condições climáticas desfavoráveis ao crescimento e produção de látex. O uso de ferramentas da biologia molecular pode ajudar na identificação de locos que controlam características quantitativas (QTLs), sobretudo a partir do desenvolvimento das tecnologias de sequenciamento de nova geração (NGS). Foi desenvolvido na presente tese um mapa genético para uma população F1 segregante (GT1 e RRIM701), onde medidas fenotípicas de diâmetro foram obtidas em diferentes idades de crescimento (12 meses, 17 meses, 22 meses, 28 meses e 35 meses), para a detecção de QTLs relacionados ao crescimento e vigor das plantas. A partir da abordagem de genotipagem por sequenciamento (GBS), foi possível selecionar 2.998 SNPs utilizando diferentes parâmetros de filtragem. Um total de 1.082 marcadores SSRs e SNPs foram mapeados, sendo 671 SNPs oriundos de GBS. Os marcadores foram distribuídos em 18 grupos de ligação e cobriram um total de 3.779,7 cM, com uma densidade de 1 marcador a cada 3,49 cM. A genotipagem de marcadores SNP através da técnica de GBS contribuiu para o aumento da densidade de marcadores, em especial em regiões onde se observava uma quantidade reduzida de marcadores SSRs. Foram detectados 19 QTLs relacionados ao diâmetro de caule em cinco épocas de crescimento avaliadas durante as estações de verão e inverno. Também foi possível observar efeitos aditivos e de dominância dos QTLs com diversos padrões de segregação auxiliando no maior entendimento da arquitetura genética de caracteres quantitativos, possibilitando novo direcionamento para o melhoramento genético da seringueira.

Abstract

The symbolic aspect that rubber tree suggests by the natural rubber production in Brazil is currently limited, mainly due the lack of resistant clones to the South America Leaf Blight disease, caused by the fungus *Microcyclus ulei*. Clones adapted to the so-called “escape regions”, where epidemic levels of disease is not observed, may be an alternative to overcome problems related to unfavorable climatic conditions for growth and latex production. Molecular biology tools usage can help to identify loci controlling quantitative traits (QTL), especially with the development of next generation sequencing technologies. It was developed in this thesis a genetic map for a full-sib population (GT1 and RRIM701) with phenotypic measures for diameter made at different growth ages (12 months, 17 months, 22 months, 28 months and 35 months) to detect QTL related to plant-growth and vigor. From the genotyping by sequencing approach (GBS), it was possible to detect 2,998 SNPs using different filtering parameters. A total of 1,082 SSR and SNP markers were mapped, with 671 SNPs from GBS. The markers were distributed into 18 linkage groups and covered a total of 3,779.7 cM with a marker density of 3.49 cM. SNP genotyping using the GBS contributed to increase marker density, especially in regions where it was observed a reduced quantity of SSR markers. We detected 19 QTL related to stem diameter at five growth periods evaluated during summer and winter seasons. It was also observed additive and dominance effects for QTL with different patterns of segregation that helps better understand the genetic architecture of quantitative traits, thus enabling new directions to genetic breeding of rubber tree.

Sumário

Organização da Tese	12
Introdução Geral.....	14
Objetivos.....	17
Objetivo Geral	17
Objetivos específicos.....	17
Revisão Bibliográfica	18
Contexto e a cultura de seringueira	18
Estudos genéticos e genômicos em seringueira	21
Marcadores moleculares e mapeamento genético em seringueira	22
Mapeamento de QTLs.....	27
Capítulo I	29
Resumo	31
Introdução	31
Material e Métodos	33
População de mapeamento e extração de DNA	33
Análise de marcadores moleculares microsatélites	34
Análise de marcadores moleculares SNPs	34
Abordagem de GBS	35
Mapeamento Genético	36
Análise Fenotípica	36
Mapeamento de QTLs	37
Resultados e Discussão	38
População de mapeamento	38
Filtragem dos marcadores e análise do polimorfismo	38
Mapeamento Genético	41
Análise Fenotípica	48
Mapeamento de QTLs	50
Agradecimentos	53
Referências do Capítulo I	54
Conclusões	60
Perspectivas	61
Referências Gerais	62
ANEXOS.....	72

Organização da Tese

Uma vez que estratégias recentes envolvendo as técnicas da biologia molecular têm provado serem úteis para diversas culturas, a presente tese buscou utilizar o polimorfismo dos marcadores moleculares para melhor entender a complexa organização genômica de seringueira (*Hevea brasiliensis*). Associado ao recente início da utilização da espécie para obtenção da borracha natural e às limitações para seu cultivo e extração de látex natural, tais metodologias são importantes para o avanço no melhoramento genético da espécie.

A partir de uma parceria da Universidade Estadual de Campinas com o Instituto Agrônomo de Campinas, foi possível desenvolver o trabalho apresentado no capítulo I, o qual descreve a utilização de marcadores SSRs (*Simple Sequence Repeats*) e SNPs (*Single Nucleotide Polymorphism*) para a construção de um mapa genético de uma população F1 segregante entre os clones GT1 e RRIM701. A robustez na identificação de marcadores SNPs por meio da abordagem de genotipagem por sequenciamento (GBS), pode contribuir para a formação do mapa genético, se destacando ainda, a utilidade dos marcadores SSRs, distribuídos ao longo de 18 grupos de ligação. Medidas fenotípicas de diâmetro do caule, tomadas em diferentes estágios de crescimento relacionados a estações de verão e inverno entre os anos de 2013 e 2015, foram utilizadas para detecção de QTLs e identificação de efeitos aditivos e de dominância, padrões de segregação e proporção da variação fenotípica explicada pelos QTLs.

Em anexo se pode encontrar trabalhos desenvolvidos em paralelo durante o período de doutorado. O trabalho no Anexo I, sob o título “*Genetic diversity strategy for the management and use of rubber genetic resources: more than 1,000 wild and cultivated accessions in a 100-genotype core collection*”, trata da análise da diversidade genética de mais de 1.000 acessos presentes em bancos de germoplasma na América do Sul, onde se observou elevados níveis de diversidade genética, bem como uma discreta separação de espécies do gênero *Hevea* a partir do conjunto de acessos totais presentes na análise de estruturação genética. Tais informações vêm a ser úteis para o melhoramento genético da espécie.

No Anexo II, se encontra o artigo intitulado “*Development of single nucleotide polymorphism markers in the large and complex rubber tree genome using next-generation sequencing*”. Devido à importância na identificação de regiões possivelmente envolvidas na regulação de promotores, marcadores SNPs foram desenvolvidos. Estes marcadores foram aproveitados na construção do mapa genético para a espécie no capítulo I.

Introdução Geral

Originária da região amazônica, *Hevea brasiliensis* (Willd. ex Adr. de Juss.) Muell-Arg, popularmente conhecida como seringueira, é considerada a maior fonte de borracha natural no mundo, a qual é extraída a partir de células especializadas conhecidas como vasos laticíferos, fornecendo borracha em quantidade e qualidade (GONÇALVES; FONTES, 2009). Matéria prima estratégica para mais de 40 mil produtos, a borracha natural tem grande importância no cenário econômico mundial, com sua produção sustentada por países do sudeste asiático (MOOIBROEK; CORNISH, 2000).

No Brasil, a seleção de clones e o plantio de seringueira é afetado pela ocorrência da doença chamada popularmente de mal-das-folhas, causada pelo fungo *Microcyclus ulei* (P. Henn. v. Arx.) (PRIYADARSHAN; CLÉMENT-DEMANGE, 2004). Apesar de ser a região de origem da seringueira, a Amazônia também fornece condições térmicas e hídricas ótimas para o desenvolvimento do fungo, o qual pode atingir toda a folhagem das seringueiras e gerar a queda prematura das folhas (CAMARGO; CAMARGO, 2008).

Estudos sobre a resistência ao fungo ainda estão em desenvolvimento (CARDOSO et al., 2014; SILVA et al., 2014b), de forma que a identificação de regiões não favoráveis a proliferação da doença, denominadas áreas de escape, se torna uma opção importante para o plantio (JAIMES et al., 2016), principalmente devido à grande demanda por borracha natural (PRIYADARSHAN; GONCALVES, 2003). Entretanto, a partir da identificação de regiões onde o cultivo não é afetado pela doença, avaliações dos efeitos climáticos, como frio e seca, merecem atenção especial para obtenção de clones adaptados (PRIYADARSHAN; GONCALVES, 2003). Atualmente, a maior parte da produção de borracha natural no Brasil é oriunda dos estados do sudeste, em especial São Paulo, onde manifestações severas da doença não são observadas e boas taxas de crescimento e produção de látex são obtidas (GONÇALVES et al., 2011; PINHEIRO; PINHEIRO, 2008).

O ciclo de melhoramento para a obtenção de um clone melhorado de seringueira pode chegar a 30 anos, isso faz com que o uso das ferramentas da biologia molecular passe a ser bastante atrativas (PRIYADARSHAN; CLÉMENT-DEMANGE, 2004), pois através destas, é possível detectar variantes alélicas

relacionadas a fenótipos de interesse e assim, levar a um aumento da eficiência e precisão no melhoramento vegetal (COLLARD; MACKILL, 2008).

Dentre as principais técnicas de biologia molecular, o uso de marcadores moleculares se tornou frequente (SCHLÖTTERER, 2004), podendo-se destacar os microssatélites e SNPs (*Single Nucleotide Polymorphism*). Associado a isso, o avanço das tecnologias de sequenciamento de nova geração (NGS) têm permitido o desenvolvimento de grande quantidade de marcadores. O método de GBS (genotipagem por sequenciamento) (ELSHIRE et al., 2011), tem se destacado, pois associa a identificação de grande quantidade de marcadores SNPs, simplicidade técnica e disponibilidade pública de *pipelines* (DE DONATO et al., 2013).

Uma das grandes aplicações dos marcadores moleculares é a construção de mapas genéticos, que além de orientar, ordenar e verificar inconsistências oriundas da montagem de sequências genômicas (FIERST, 2015), também podem auxiliar na seleção de plantas que apresentem fenótipos de interesse, pois permite identificar marcadores ligados a genes e regiões que controlam características quantitativas (LE GUEN et al., 2011a; SOUZA et al., 2013).

Neste contexto, e na busca pelo melhor entendimento da complexa arquitetura genética de *Hevea brasiliensis*, foi realizado, por meio de marcadores SNPs e microssatélites, o desenvolvimento de um mapa genético para a espécie, bem como o mapeamento de QTLs (*Quantitative Trait Loci*) associados a características relacionadas ao crescimento em diferentes épocas do ano.

A perspectiva de se utilizar os marcadores moleculares para identificação de genes e QTLs visando o aumento da eficiência e precisão do melhoramento convencional de *Hevea brasiliensis*, tem se tornado cada vez mais possível, pois considerações como custo operacional, nível de polimorfismo, procedimentos técnicos e magnitude da detecção de novos marcadores, já não são um problema para a espécie. O panorama atual reflete a necessidade de continuar explorando os efeitos e posições de QTLs e, posteriormente, aplicar as descobertas através de seleção assistida por marcadores.

O desenvolvimento deste trabalho foi possível por meio da parceria entre a Universidade Estadual de Campinas (UNICAMP), o Centro Avançado de Pesquisa da Borracha/IAC, a Escola Superior de Agricultura “Luiz de Queiroz” - USP e o Centre de

Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD).

Objetivos

Objetivo Geral

- Desenvolvimento de novos marcadores SNPs e construção de mapa genético de *Hevea brasiliensis* para detecção de QTLs associados a características fenotípicas de interesse no melhoramento genético da espécie.

Objetivos específicos

- Desenvolvimento de marcadores SNPs a partir de tecnologias de sequenciamento de nova geração.
- Construir um mapa genético saturado de *Hevea brasiliensis*.
- Detectar QTLs associados com crescimento vegetal em épocas de verão e inverno.
- Analisar os efeitos de possíveis QTLs sobre o diâmetro do caule em diferentes condições climáticas.

Revisão Bibliográfica

Contexto e a cultura de seringueira

Hevea brasiliensis (Figura 1), popularmente conhecida como seringueira, é uma espécie arbórea originária da região amazônica, pertencente à família Euphorbiaceae, que inclui outros importantes gêneros de culturas tropicais, tais como *Ricinus* (mamona) e *Manihot* (mandioca). O gênero *Hevea* por sua vez, apresenta 11 espécies e se trata de um táxon bem definido e de fácil reconhecimento, o que não se pode dizer das suas espécies, que são dificilmente separáveis uma das outras (PIRES; SECCO; GOMES, 2002) e apresentam frequente hibridação e introgressão (GONÇALVES; FONTES, 2009).



Figura 1: *Hevea brasiliensis*: a) indivíduo com porte arbóreo podendo atingir 30 metros de altura; b) folhas; c) sangria de látex.

A ocorrência natural do gênero *Hevea*, abrange toda a extensão da bacia amazônica, incluindo parte do Brasil, Bolívia, Peru, Colômbia, Equador, Venezuela, Guiana Francesa, Suriname e Guina (GONÇALVES; FONTES, 2009). A carta de clima de Köppen mostra nessas regiões uma temperatura mínima $\geq 18^{\circ}\text{C}$ e umidade constante ou com curta estação de seca (PEEL; FINLAYSON; MCMAHON, 2007).

Pertencente a classe das dicotiledôneas, o gênero é descrito como monóico e alógamo, com flores unissexuadas amarelas dispostas em racimo. Suas folhas são

longamente pecioladas, repartidas em três folíolos e seu fruto é do tipo cápsula, geralmente com três sementes (PAIVA, 1992). Todas as espécies, exceto *Hevea spruceana* e *Hevea microphylla*, possuem deiscência e apresentam látex em todas as partes da planta (GONÇALVES; MARQUES, 2008). A polinização é realizada pelo vento e por pequenos insetos da família Ceratopogonidae (Heleidae), operando em curtas distâncias (IPEF, 2007). Quanto a dispersão de sementes, Le Guen et al. (2009) levanta a possibilidade de ser realizada através de alagamentos sazonais da bacia amazônica, já que não há possibilidade de dispersão pelo vento devido ao seu peso e tamanho e a disseminação por meio de pássaros e mamíferos não ultrapassar algumas centenas de metros.

Podendo atingir 30 metros de altura e 1,5 metros de diâmetro do caule sob condições favoráveis, *Hevea brasiliensis* trata-se de uma planta de hábito decíduo. Em regiões da Amazônia, onde ou quando períodos secos são menos intensos, a queda de folhas e o florescimento são irregulares. Por sua vez, quando cultivadas no estado de São Paulo, normalmente a senescência ocorre no período de julho a agosto (GONÇALVES; MARQUES, 2008).

Considerada como matéria prima estratégica para mais de 40 mil produtos, a borracha natural no gênero *Hevea* é extraída a partir de células especializadas conhecidas como vasos laticíferos. Maior fonte de borracha natural do mundo, *Hevea brasiliensis* é a espécie mais importante do gênero e a única capaz de suprir as necessidades do mercado em quantidade e qualidade (GONÇALVES; FONTES, 2009).

Foi a partir de 1838, com a descoberta do processo de vulcanização, por Charles Goodyear, e crescimento do interesse pela borracha natural, que a espécie começou a ser valorizada (MOOIBROEK; CORNISH, 2000). Com a introdução de um pequeno número de mudas no oriente em 1876 por Henry Wickham, se pode realizar a seleção de clones produtivos, e assim, tornar os países do sudeste asiático os maiores produtores de borracha natural no mundo (SECCO, 2008).

A influência da borracha no Brasil, chegou a caracterizar uma época denominada “ciclo da borracha” (GONÇALVES; CARDOSO; ORTOLANI, 1990), entretanto, a produção de borracha natural atualmente é limitada pelo fungo *Microcyclus ulei* (P. Henn. v. Arx.), responsável pela doença conhecida como mal-

das-folhas (CARDOSO et al., 2014). A grande susceptibilidade da seringueira ao fungo pode ser observada a partir de seis meses após o plantio (RIVANO et al., 2010), de forma que condições térmicas e hídricas ótimas para o desenvolvimento do fungo encontradas na Amazônia, resultam em grande quantidade de esporos, desenvolvimento de epidemias, e perecimento do seringal (CAMARGO; CAMARGO, 2008). Por sua vez, grandes países asiáticos produtores de borracha não registraram até o momento sinais da doença apesar das condições climáticas favoráveis para severa infecção, possivelmente explicado pelo intenso controle de tráfego internacional para tais países (LIEBEREI, 2007).

Diversas prospecções foram realizadas na Amazônia no século XX com intuito de se obter recursos genéticos de *Hevea brasiliensis*. Tais recursos, armazenados em bancos de germoplasma, podem auxiliar melhoristas na tentativa de se encontrar a resistência natural para o mal-das-folhas, entretanto, poucos estudos têm sido realizados até o momento. Souza et al. (2015) recentemente descreveram pela primeira vez a diversidade genética de mais de 1.000 acessos de *Hevea* mantidos nos principais bancos de germoplasma da América do Sul (trabalho em anexo).

A elevada capacidade do fungo *Microcyclus ulei* desenvolver novas cepas e superar a resistência selecionada (LE GUEN et al., 2008), reflete na recomendação até o momento, apenas de genótipos parcialmente resistentes (CARDOSO et al., 2014; SILVA et al., 2014b), além disso, destaca-se a base genética bastante reduzida dos materiais cultivados (SOUZA et al., 2015). Dessa forma, a expansão do cultivo para regiões onde as condições climáticas permitam o crescimento das plantas em detrimento da ocorrência de níveis epidêmicos do mal das folhas, denominadas áreas de escape, passa a ser uma alternativa importante para a heveicultura (PRIYADARSHAN; GONCALVES, 2003).

Uma importante área de cultivo de seringueira são os estados do sudeste do Brasil, em especial São Paulo, onde, apesar do período frio e seco característico durante o inverno, se observa boas taxas de crescimento e produção de látex (PINHEIRO; PINHEIRO, 2008). Gonçalves et al. (2011) ao avaliar a produção de borracha em clones das séries IAC 500 durante 12 anos, plantados na região noroeste do estado de São Paulo, obteve valores médios de até 1.731 kg de borracha por hectare por ano. O potencial dessa região é evidenciado devido às condições

climáticas não favoráveis para a ocorrência do mal-das-folhas em níveis epidêmicos (CENTURION et al., 2005).

O melhoramento genético de seringueira vem buscando clones adaptados a áreas de escape como o estado de São Paulo. Dentre os objetivos, destaca-se a obtenção de clones com alto potencial de produção, e outros caracteres secundários desejáveis, que contribuem para o aumento da produtividade, tais como vigor, crescimento, tolerância à quebra pelo vento, tolerância à frio, dentre outros. (GONÇALVES; MARQUES, 2008).

Estudos genéticos e genômicos em seringueira

Em *Hevea*, o número de cromossomos é caracterizado como $2n = 36$ (RAMAER, 1935). É descrito que a seringueira se comporta como um diplóide, apesar de parecer ser um anfidiplóide ($2n = 4x = 36$) que se estabilizou durante o curso da evolução (PRIYADARSHAN; GONCALVES, 2003). Leitch et al. (1998) interessados em entender os efeitos da ausência de barreiras genéticas e consequente ocorrência de híbridos naturais para o gênero (WYCHERLEY, 1976), bem como a presença de constrições secundárias em quatro cromossomos indicando quatro pares de locos de rDNA (ONG, 1975), sugerem uma origem alotetraplóide para o gênero após observar apenas um loco para o rDNA 5S e dois locos para o rDNA 18S-25S em cromossomos distintos.

Trabalhos iniciais de biologia molecular em *Hevea brasiliensis* são notados a partir de clonagens e caracterização de genes envolvidos na biossíntese de látex nos anos de 1990's (SAHA; PRIYADARSHAN, 2012). Adicionalmente, progressos no entendimento da genética e genômica de seringueira utilizando diversas metodologias puderam ser realizadas (CHOW et al., 2012, 2007; LE GUEN et al., 2011b; SILVA et al., 2014a; SOUZA et al., 2009). A grande quantidade de dados gerados pelas tecnologias de sequenciamento de nova geração (LAU et al., 2016; MANTELLO et al., 2014; RAHMAN et al., 2013; SOUZA et al., 2016; TANG et al., 2016; XIA et al., 2011), possibilitaram ainda mais investigações, de forma que foi possível identificar, por exemplo, a ocorrência de rearranjos estruturais, envolvidos na fusão de transcritos relacionados à síntese de ATP nas mitocôndrias e consequente geração de energia insuficiente para o desenvolvimento de anteras, explicando a herança materna da macho-esterilidade em alguns clones, como GT1 e BPM24 (SHEARMAN et al., 2014).

Estudos genômicos em seringueira, entretanto, ainda estão em seu início (SAHA; PRIYADARSHAN, 2012). A recente intensificação de trabalhos para obtenção de melhores montagem das sequências genômicas, buscando melhor conhecer detalhes moleculares da biossíntese de borracha (LAU et al., 2016; RAHMAN et al., 2013; TANG et al., 2016), tem ajudado no estabelecimento de perspectivas em estudos de associação genômica, melhoramento de germoplasma e seleção assistida por marcadores, porém, tais trabalhos evidenciam os desafios na obtenção de genoma bem montados de *Hevea brasiliensis*, principalmente devido à grande quantidade de sequências repetitivas (>70%) e elevada heterozigosidade (LAU et al., 2016).

Novas considerações a respeito do tamanho estimado do complexo genoma de seringueira realizadas com o auxílio de bibliotecas de BACs (TANG et al., 2016) e da plataforma PacBio (Pacific Biosciences, Menlo Park, CA, USA) (LAU et al., 2016) sugerem um genoma com 1,47Gb e 1,55Gb, respectivamente. Análises comparativas entre *Hevea brasiliensis* e outras espécies revelam forte sintenia entre as sequências genômicas com membros da família Euphorbiaceae como *Manihot esculenta*, *Ricinus communis* e *Jatropha curcas*, embora grande variação no tamanho do genoma possa ser encontrado, como no caso de *M. esculenta*, com genoma estimado em 742Mb (WANG et al., 2014) e *R. communis* com 350Mb (CHAN et al., 2010).

Marcadores moleculares e mapeamento genético em seringueira

Os marcadores moleculares vêm se tornando uma ferramenta importante para o melhoramento vegetal tanto na abordagem clássica quanto na moderna (YANG et al., 2015). Por sua vez, marcadores moleculares revelam polimorfismos entre indivíduos em nível de DNA, possibilitando a obtenção de informações sobre a organização da diversidade genética e também, a determinação de regiões do genoma que contribuem para a variação de diversas características de importância agrônoma (CHARCOSSET; MOREAU, 2004).

Com o advento das técnicas modernas de biologia molecular, como a PCR (*Polymerase Chain Reaction*) e a tecnologia de sequenciamento de nova geração (NGS), o uso de marcadores moleculares se tornou frequente (SCHLÖTTERER, 2004; YANG et al., 2015). Desde então, avanços significativos nos estudos genéticos têm sido obtidos, principalmente no entendimento de genomas complexos, gerando

novas oportunidades para a genética e melhoramento vegetal (EDWARDS; BATLEY; SNOWDON, 2013).

Uma das classes de marcadores moleculares mais populares nos estudos genéticos são os marcadores do tipo microssatélites ou SSRs (*Simple Sequence Repeats*), os quais são sequências de 1 a 6 pares de base repetidas em tandem (ZANE; BARGELLONI; PATARNELLO, 2002). Caracterizados por apresentarem herança mendeliana codominante, são também considerados multialélicos, apresentam regiões flangeadoras conservadas e possuem alta reprodutibilidade (FERREIRA; GRATTAPAGLIA, 1996).

O polimorfismo dos marcadores microssatélites advém da diversidade do número de motivos repetidos, fornecendo um elevado nível de informação genética. Dentre a grande quantidade de aplicações em genética vegetal, tais marcadores têm sido utilizados em trabalhos envolvendo mapeamento genético (MCCOUCH et al., 1997; OLIVEIRA et al., 2007; YANG et al., 2016), genética de populações (CONSON et al., 2013; MORI; ZUCCHI; SOUZA, 2015; SANTOS; CERQUEIRA-SILVA; MORI, 2015) e seleção assistida por marcadores (BENCHIMOL; SOUZA JR; SOUZA, 2005).

Por outro lado, devido ao desenvolvimento das tecnologias de sequenciamento de nova geração, os marcadores SNPs (*Single Nucleotide Polymorphism*) têm se destacado na análise genética moderna, já que a produção de dados não mais é um fator limitante (LEE et al., 2011). Extremamente abundantes no genoma, marcadores SNPs são variações de uma única base entre sequências de DNA de indivíduos em uma população, podendo ser representado como transversões (C/G, A/T, C/A e T/G), transições (C/T ou G/A) e inserções/deleções (InDels). A maioria dos SNPs são bi-alélicos, embora também exista aqueles classificados como tri-alélicos e tetra-alélicos (HAYWARD et al., 2012).

Baseando-se no fato de um SNP resultar na tradução de um mesmo aminoácido ou não, estes podem ser classificados como sendo marcadores sinônimos ou não-sinônimos, respectivamente. Para o caso de SNPs não-sinônimos que ocorrem em regiões codantes, mudanças no fenótipo podem ser observadas, pois tais eventos estão relacionados a alterações na estrutura ou função da proteína e assim o indivíduo passa a apresentar respostas diferentes às pressões do ambiente (HAYWARD et al., 2012).

Diversos métodos de genotipagem de marcadores SNPs, como por exemplo a plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, CA, USA), a genotipagem por sequenciamento (GBS) (ELSHIRE et al., 2011), RAD-seq (BAIRD et al., 2008), CRoPS (VAN ORSOUW et al., 2007), têm sido utilizados em diferentes espécies. Os avanços em tais metodologias está relacionado à maneira robusta para a obtenção de dados.

Envolvendo a digestão de DNA genômico com enzimas de restrição para reduzir a complexidade do genoma, a abordagem de GBS permite a identificação e genotipagem ao mesmo tempo, de grande quantidade de SNPs. Para isso, são utilizados adaptadores contendo *barcodes* e adaptadores comuns em uma reação de ligação, para produzir bibliotecas multiplexadas de amostras a serem sequenciadas (POLAND; RIFE, 2012). Assim, fragmentos de DNA (*tags*) amplificados com sucesso e específicos a cada indivíduo, podem posteriormente, ser selecionados para a rápida descoberta de SNPs (DAVEY et al., 2011).

Tais como em outros marcadores mais antigos como os AFLPs (VOS et al., 1995), ou mesmo os marcadores de RAD-seq (BAIRD et al., 2008), o uso de enzimas de restrição visa produzir uma representação do genoma (DAVEY et al., 2011). Por meio da escolha da enzima de restrição apropriada, regiões repetitivas do genoma podem ser evitadas e regiões de baixo número de cópias podem ser amostradas com duas a três vezes mais eficiência, de maneira a simplificar desafios no alinhamento das sequências em espécies com altos níveis de diversidade genética (ELSHIRE et al., 2011). Esta abordagem demonstrou ser eficaz em muitas espécies, e foi capaz de produzir dezenas de milhares de marcadores SNPs (ELSHIRE et al., 2011; LU et al., 2013).

Baseado na plataforma de sequenciamento da Illumina (Illumina, San Diego, CA, USA), e sendo originalmente pensado para espécies vegetais (ELSHIRE et al., 2011), o método de GBS (Figura 2) tem aplicações em estudos populacionais, caracterização de germoplasma, melhoramento genético e saturação de mapas de ligação, para uma grande variedade de organismos, sendo as maiores vantagens sobre outros protocolos, relacionadas à simplicidade técnica e disponibilidade pública de pipelines (DE DONATO et al., 2013). Por outro lado, o grande número de dados gerados pela metodologia, necessita de cuidado especial, como a criação de algoritmos de imputação e estabelecimento de parâmetros de filtragem para lidar com

grande quantidade de dados perdidos e genótipos falso positivo gerados, tais como realizado em *Rubus idaeus* (WARD et al., 2013).

A escolha de qual marcador utilizar em pesquisas de genética vegetal está intimamente relacionada a questionamentos como custo, tamanho amostral, densidade de marcadores, recursos genômicos disponíveis, de forma que a associação de marcadores SSRs e SNPs pode se tornar uma abordagem valiosa (HODEL et al., 2016).

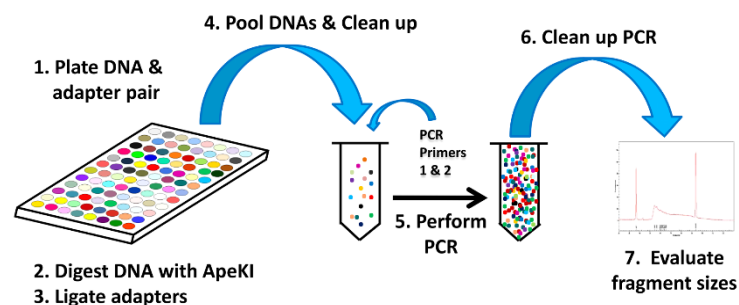


Figura 2: Esquema da técnica de genotipagem por sequenciamento (GBS).

Em um cenário onde a tecnologia de sequenciamento de nova geração tem revolucionado as análises genéticas, revelando novas oportunidades para o melhoramento genético de diversas culturas, o uso de marcadores moleculares para mapeamento genético continua importante, já que este, permite verificar inconsistências no mapa físico. Também, a associação entre mapas genéticos e projetos de sequenciamento, não só podem ajudar no melhor entendimento da organização genômica (FIERST, 2015), mas também pode auxiliar na detecção de genes ou QTLs que controlam características de importância econômica (COLLARD; MACKILL, 2008).

Baseado no princípio de recombinação genética durante a meiose (TANKSLEY, 1993), mapas genéticos são formados por grupos de ligação nos quais está representada a ordem e a distância entre marcadores genéticos. Considerações como o estabelecimento de populações segregantes adequadas, tipos e quantidade de marcadores a serem utilizados, características de interesse econômico presente nos genitores e metodologias genético-estatísticas apropriadas devem ser feitas antes da realização de cada etapa da construção de mapas genéticos (CARNEIRO; VIEIRA, 2002).

A construção de mapas genéticos para espécies dentre as quais a obtenção de linhagens endogâmicas não é possível, como por exemplo em espécies florestais, é realizado a partir de um cruzamento entre genitores não homozigóticos, formando uma população de irmãos completos (MARGARIDO; SOUZA; GARCIA, 2007). Isso é possível pois Wu et al. (2002) propuseram um método baseado em máxima verossimilhança que permite a estimação simultânea da ligação e fases de ligação entre marcadores e assim a utilização de todas as classes de marcadores informativos.

Vários mapas genéticos já foram construídos para *Hevea brasiliensis* até o momento utilizando-se de diferentes metodologias (LE GUEN et al., 2011a; LESPINASSE et al., 2000; NIRAPATHPONGPORN et al., 2016; NOVALINA, 2013; POOTAKHAM et al., 2015; SHEARMAN et al., 2015; SOUZA et al., 2013; TRIWITAYAKORN et al., 2011), sendo possível verificar que o desenvolvimento de mapas genéticos para a espécie sempre seguiu estratégias disponíveis à época na tentativa de se obter avanços no entendimento de sua arquitetura genética.

Os primeiros mapas para a espécie (LE GUEN et al., 2011a; LESPINASSE et al., 2000; NOVALINA, 2013; TRIWITAYAKORN et al., 2011), por meio de marcadores RFLPs (*Restriction Fragment Length Polymorphism*), RAPDs (*Randon Amplified Polymorphism DNA*), AFLPs (*Amplified Fragment Length Polymorphism*), SSRs e isoenzimas, utilizaram uma estratégia chamada duplo *pseudo-testcross* (GRATTAPAGLIA; SEDEROFF, 1994) onde, mapas individuais para cada genitor eram produzidos através da identificação de marcadores polimórficos segregando na proporção 1:1. Por sua vez, Souza et al. (2013) usando 284 marcadores SSRs, dos quais 146 classificados como mais informativos (com segregação 1:1:1:1 e 1:2:1) segundo Wu et al. (2002), construíram um mapa genético integrado, representando mais precisamente os polimorfismos observados nos genitores. Recentemente, o desenvolvimento e genotipagem de enorme quantidade de SNPs oriundos de NGS foi realizado para *Hevea brasiliensis* (POOTAKHAM et al., 2015; SHEARMAN et al., 2015) na tentativa de se obter mapas genéticos saturados baseados exclusivamente nessa classe de marcadores.

Apesar da realização do mapeamento genético utilizando dados de NGS ser robusta devido a rápida obtenção de grande quantidade de marcadores (YANG et al., 2015), as metodologias para construção de mapas genéticos foram originalmente

desenvolvidas para dados em pequena escala e assim, a divisão dos grupos de ligação e a ordenação e cálculo da distância entre marcadores tende a ser um desafio computacional. Destaca-se ainda nessas plataformas, um aumento na quantidade de erros na identificação dos alelos e grande quantidade de dados perdidos oriundos das limitações dessas tecnologias. Dois ou mais SNPs podem, por exemplo, ser artificialmente colapsados em um único marcador devido à similaridade de sequências em repetições, baixa complexidade de regiões e presença em famílias gênicas duplicadas, representando informações redundantes, as quais devem ser consideradas com cuidado (FIERST, 2015).

A construção de mapas genéticos utilizando-se inicialmente marcadores microssatélites como uma estrutura de suporte essencial aos grupos de ligação de espécies com genoma complexo, é uma ferramenta valiosa na perspectiva de uso de marcadores oriundos de NGS (YANG et al., 2016). Dessa forma, os diversos trabalhos envolvendo o desenvolvimento de marcadores microssatélites em *Hevea brasiliensis* (FENG et al., 2009; LE GUEN et al., 2011b; MANTELLO et al., 2012; SILVA et al., 2014a; SOUZA et al., 2009), se revelam como fontes valiosas de informação.

Mapeamento de QTLs

O entendimento da conexão entre genótipo e fenótipo para caracteres complexos tem provado ser desafiador e, muito embora a genômica tenha fornecido novas possibilidades, investigações rigorosas necessitam ser realizadas. A essas regiões genômicas complexas que controlam caracteres quantitativos, denomina-se QTL ou *Quantitative Trait Loci* (DOERGE, 2002).

Estudos têm mostrado grande quantidade de marcadores moleculares ligados a QTLs (BRADSHAW et al., 2008; COLLARD; MACKILL, 2008; KING et al., 2015; LI; HE, 2014; ZUREK; TOPP; BENFEY, 2015), de forma que a detecção dessas regiões complexas que controlam características de importância econômica com o auxílio de marcadores e análise genética de ligação (TANKSLEY, 1993), passa a ter grande importância no melhoramento vegetal, ainda mais com a grande disponibilidade de marcadores oriundos das tecnologias de NGS.

Dentre os modelos gerais de análise e mapeamento de QTLs, nota-se: (i) a análise de marcas simples, que é útil quando o objetivo é somente testar se há alguma associação entre fenótipo e genótipo; (ii) o mapeamento por intervalo simples,

proposto por Lander e Botstein, (1989), que considera marcadores adjacentes e propicia estimativas da localização genética e efeitos dos QTLs; (iii) o mapeamento por intervalo composto ou CIM (JANSEN; STAM, 1994; ZENG, 1994), que considera os efeitos de marcadores fora do intervalo estudado e leva a melhoras significativas na localização de QTLs; (iv) mapeamento por intervalos múltiplos (KAO; ZENG; TEASDALE, 1999), que considera múltiplos intervalos simultaneamente para a detecção de QTLs e permite incluir os efeitos epistáticos no modelo.

Desde o seu desenvolvimento, o modelo de mapeamento por intervalo composto tem sido amplamente utilizado para populações formadas a partir de linhagens, sendo Gazaffi et al. (2014) responsável pela extensão da metodologia para espécies as quais apenas se pode obter populações a partir de cruzamentos entre indivíduos não-endogâmicos. Com esse modelo, estimativas da probabilidade dos QTLs são obtidas através de análise multiponto via modelo de cadeia de Markov, permitindo o uso de marcadores moleculares com diversos padrões de segregação. Tal metodologia foi utilizada por Souza et al. (2013) na detecção de 18 QTLs de *Hevea brasiliensis* para características de crescimento em diferentes estações do ano. O uso desse método é importante, pois ajuda na identificação de marcadores genéticos que estão ligados próximos a genes de importância econômica, bem como identificação da origem dos alelos que levam a alterações no fenótipo e assim, tornando-se possível a aplicação da seleção assistida por marcadores (GAZAFFI et al., 2014).

Capítulo I

Mapeamento genético-molecular e análise de QTLs associados ao crescimento em *Hevea brasiliensis* utilizando marcadores microssatélites e SNPs

André Ricardo Oliveira Conson¹, Isabela Aparecida Araújo Andreotti¹, Cristiane Hayumi Taniguti², Rodrigo Rampazo Amadeu², João Ricardo Bachega Feijó Rosa², Luciano Henrique Brás dos Santos¹, Carla Cristina da Silva¹, Camila Campos Mantelo¹, Erivaldo José Scaloppi Junior³, Paulo de Souza Gonçalves³, Vincent Le Guen⁴, Antônio Augusto Franco Garcia², Lívia Moura de Souza¹, Anete Pereira de Souza^{1,5}

¹ Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, Brasil

² Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz/USP, Piracicaba, Brasil

³ Centro Avançado de Pesquisa da Borracha, Instituto Agrônomo de Campinas, Votuporanga, Brasil

⁴ Centre de Coopération Internationale em Recherche Agronomique pour le Développement (CIRAD), UMR AGAP, Montpellier, França

⁵ Departamento de Biologia Vegetal – Instituto de Biologia, Universidade Estadual de Campinas, Campinas, Brasil

Resumo

Originária da região amazônica, *Hevea brasiliensis*, popularmente conhecida como seringueira, tem grande importância devido a elevada produção de borracha natural. No Brasil, o plantio de seringueira é restrito a regiões não favoráveis à proliferação do principal fitopatógeno, chamado *Microcyclus ulei*, porém, condições climáticas desfavoráveis passam a ser um desafio ao crescimento e produção de látex. Marcadores moleculares surgem como uma excelente opção na busca por regiões genômicas que controlam características quantitativas. Assim, por meio de marcadores SNPs e SSRs, um mapa genético para uma população F1 segregante (GT1 e RRIM701) foi desenvolvido e medidas de crescimento do diâmetro obtidas em épocas de inverno e verão visando a detecção de QTLs. Um total de 1.082 marcadores, dos quais 671 SNPs de alta qualidade obtidos a partir da abordagem de genotipagem por sequenciamento (GBS), foram distribuídos em 18 grupos de ligação e cobriram um total de 3.779,7 cM, com uma densidade de marcadores de 3,49 cM. Regiões com baixa densidade de marcadores SSRs puderam ser superadas por meio da abordagem GBS. Um total de 19 QTLs relacionados ao diâmetro de caule foram detectados em diferentes grupos de ligação, com a maior parte após períodos de verão e três para o inverno. Também foi possível observar efeitos aditivos e de dominância dos QTLs com diversos padrões de segregação. Reflexões a respeito de QTLs estáveis ao longo das avaliações indicam as vantagens da obtenção de mapas com elevada densidade de marcadores e também a possibilidade de expressão de vários QTLs, ou mesmo vários genes ao longo do tempo. Informações sobre efeitos e localização de QTLs possivelmente detectados podem dar suporte para o melhoramento genético da espécie.

Palavras chave

Seringueira, *Hevea brasiliensis*, mapa genético, QTL, GBS

Introdução

Originária da Amazônia e principal produtora de látex natural no mundo, *Hevea brasiliensis* (Willd. ex A.D. de Juss.) Muell-Arg., também conhecida como seringueira, é uma espécie perene, alógama, pertencente à família Euphorbiaceae e que apresenta hibridação com outras espécies do gênero (PRIYADARSHAN; CLÉMENT-DEMANGE, 2004). Contendo 36 cromossomos ($2n = 36$), comporta-se como um

diplóide, apesar de estudos indicarem ser um anfidiplóide ($2n = 4x = 36$) que se estabilizou durante o curso da evolução (SAHA; PRIYADARSHAN, 2012). Análises genômicas para a espécie têm se mostrado complexas, necessitando utilizar várias metodologias para um maior entendimento de sua organização (LAU et al., 2016; RAHMAN et al., 2013; SOUZA et al., 2016; TANG et al., 2016).

A descrição do plantio de seringueira em regiões chamadas “áreas de escape” como importante alternativa para a produção de borracha natural (JAIMES et al., 2016; SOUZA et al., 2013), está relacionada às limitações oriundas da ocorrência da doença mal-das-folhas, causada pelo fungo *Microcyclus ulei* (P.Henn. v. Arx.) (PRIYADARSHAN; CLÉMENT-DEMANGE, 2004), cujos estudos sobre a resistência ainda estão em desenvolvimento (CARDOSO et al., 2014; SILVA et al., 2014b). Exemplo da possibilidade de sucesso é o plantio na região sudeste do Brasil, onde manifestações severas da doença não são observadas e boas taxas de crescimento e produção de látex podem ser obtidas (GONÇALVES et al., 2011; PINHEIRO; PINHEIRO, 2008). Em tais regiões, entretanto, baixas temperaturas e índices pluviométricos reduzidos são comuns em alguns meses do ano, o que afeta o desenvolvimento da planta e, conseqüentemente, a produção de borracha (RAO; KOLE, 2016; SILPI et al., 2006).

Com o auxílio dos marcadores moleculares, é possível realizar a identificação de variantes alélicas relacionadas a fenótipos adaptados à essas regiões, de forma a encurtar o longo ciclo de melhoramento da espécie (PRIYADARSHAN; CLÉMENT-DEMANGE, 2004). Nesse sentido, resultados significativos podem ser obtidos a partir dos avanços das tecnologias de sequenciamento de nova geração (NGS) e de estratégias que permitam a descoberta e genotipagem de SNPs em múltiplos indivíduos simultaneamente, como o método de GBS (genotipagem por sequenciamento) (ELSHIRE et al., 2011).

Em *Hevea brasiliensis*, os marcadores moleculares têm sido úteis para a construção de mapas genéticos (LE GUEN et al., 2011a; LESPINASSE et al., 2000; NIRAPATHPONGPORN et al., 2016; NOVALINA, 2013; POOTAKHAM et al., 2015; SHEARMAN et al., 2015; SOUZA et al., 2013; TRIWITAYAKORN et al., 2011) e detecção de marcadores ligados a genes e QTLs (LE GUEN et al., 2011a; SOUZA et al., 2013). Com a disponibilidade de métodos de genotipagem em larga escala e de

sequências genômicas para a espécie (LAU et al., 2016; RAHMAN et al., 2013; TANG et al., 2016), QTLs adicionais podem ser localizados.

A partir de marcadores SNPs e microssatélites previamente utilizados em uma dissertação de mestrado para a construção de um mapa genético, o presente trabalho buscou desenvolver marcadores SNPs oriundos da abordagem de GBS para maior saturação do mapa genético para o cruzamento avaliado. Foi possível realizar também o mapeamento de QTLs associados a uma característica relacionada ao crescimento em diferentes épocas do ano.

Material e Métodos

População de mapeamento e extração de DNA

Para a construção do mapa genético, foi utilizado uma população F1 segregante oriunda do cruzamento entre GT1 e RRIM701 com 146 indivíduos, os quais foram obtidos a partir de polinização aberta. Originário da coleta de Wickham em 1876 no estado do Pará, GT1 é classificado como macho-estéril (SHEARMAN et al., 2014), tolerante a vento e frio, enquanto que RRIM701, oriundo do programa de melhoramento asiático, se apresenta com crescimento vigoroso e bom aumento da circunferência após o início da sangria (CIRAD, 2011).

Dez marcadores microssatélites foram utilizados para conferência da paternidade das plantas para, em seguida, prosseguir com a instalação da população de mapeamento no Centro Avançado de Pesquisa da Borracha, órgão associado ao Instituto Agrônomo de Campinas, na região noroeste do estado de São Paulo (20°25'00"S e 49°59'00"O a 450 metros de altitude) no ano de 2012 em blocos aumentados com quatro repetições. Cada bloco consistiu de aproximadamente 37 genótipos do cruzamento GT1 x RRIM701 e quatro testemunhas (RRIM600, GT1, PB235 e RRIM701).

As folhas da população de mapeamento foram liofilizadas e moídas antes da extração de DNA utilizando CTAB (*Cetyl trimethylammonium bromide*) 2% conforme metodologia descrita por Doyle e Doyle, (1987). O DNA genômico produzido, foi analisado em gel de agarose 1% para verificar sua integridade e sua concentração, estimada via QuantiFluor-ST Fluorometer (Promega, Madison, WI, USA).

Análise de marcadores moleculares microssatélites

Foram testados 1.174 marcadores microssatélites para avaliação do polimorfismo nos genitores da população de mapeamento, entre estes, 333 desenvolvidos no Laboratório de Análise Genética e Molecular da Universidade Estadual de Campinas (MANTELLO et al., 2012; SILVA et al., 2014a; SOUZA et al., 2009), 296 desenvolvidos por Le Guen et al. (2011b), 184 EST-SSR desenvolvidos por Feng et al. (2009) e 361 EST-SSR (não publicado) construídos a partir de bibliotecas de hibridização por supressão subtrativa utilizando-se folhas inoculadas com o fungo *M. ulei* (GARCIA et al., 2011). Ao todo, 287 marcadores SSRs foram selecionados baseando-se no seu nível de informação, como discutido por Wu et al. (2002).

A amplificação das regiões de microssatélites foi realizada conforme publicado pelos autores, e a visualização dos produtos de PCR, por meio de três técnicas: coloração com nitrato de prata (CRESTE; NETO; FIGUEIRA, 2001), LI-COR 4300 DNA analyzer, visualizando as bandas por meio de fluorescência, e sequenciador automático 3500XL DNA ABI (Life Technologies, Carlsbad, CA, USA). Durante as genotipagens observou-se que dois genótipos se apresentaram como sendo contaminantes e foram excluídos das análises.

Análise de marcadores moleculares SNPs

A partir de ESTs derivados de bibliotecas de cDNA (SILVA et al., 2014a) e sequências transcritas de painel de seringueira (MANTELLO et al., 2014), 391 SNPs foram escolhidos para genotipagem da população de mapeamento por meio da plataforma Sequenom iPLEX MassARRAY® (Sequenom Inc., San Diego, Califórnia, USA). Tais SNPs foram anotados com as vias de mevalonato e MEP (2-C-methyl-D-erythritol 4-phosphate), conhecidamente envolvidas na biossíntese de látex, além de vias relacionadas à síntese de madeira.

Após multiplexar o ensaio por meio do *software* Sequenom Assay Design versão 2.0, uma PCR foi realizada em um total de 5µL, sendo 2 ng de DNA genômico, 0,5 U HotStar Taq *polymerase* (QIAGEN, Chatsworth, CA), 100 nM de *primers*, 1,25x HotStar Taq *buffer*, 1,62 mM de MgCl₂, e 500 µM de dNTPs. Após a ativação da enzima a 94°C por 15 minutos, seguiu-se para amplificação do DNA com 45 ciclos de 94°C por 20 segundos, 56°C por 30 segundos e 72°C por 1 minuto, com extensão final

durante 5 minutos a 72°C. Os dNTPs não incorporados, foram eliminados por meio da enzima *shrimp alkaline phosphatase* (0,3 U; Sequenom, Inc., San Diego, CA). Uma segunda PCR com função de adição de uma base única foi realizada por meio da adição de *primers* de extensão em concentrações de 0,625 uM (*primers* com baixo peso molecular) a 1,25 uM (*primers* com alto peso molecular). Enzima iPLEX e tampões também foram adicionados na reação com volume total de 9 µL. Por fim, as reações foram submetidas ao MassARRAY Compact System e os espectros de massa analisados usando o *software* MassArray Typer 4.0 Analyzer (Sequenom, Inc., San Diego, CA), para visualização de genótipos e frequências alélicas. Desses, 159 marcadores SNPs foram identificados com segregação 1:1 ou 1:2:1 na população de mapeamento e puderam ser incluídos nas análises do mapa genético.

Outros 96 SNPs polimórficos obtidos a partir de sequenciamento de amostras de DNA genômico de folhas de *Hevea brasiliensis* (SOUZA et al., 2016 - em anexo) e de dados de transcriptoma foram utilizados para carregar um *chip* 96.96 da plataforma BioMark Real-Time PCR System (Fluidigm, Inglaterra) e assim, genotipagem da população de mapeamento. Tal metodologia foi desenvolvida na dissertação de mestrado da aluna Isabela Aparecida Araújo Andreotti.

Abordagem de GBS

A genotipagem por sequenciamento (GBS) foi realizada no Laboratório de Análise Genética e Molecular da Universidade Estadual de Campinas, tomando por base o protocolo desenvolvido por Elshire et al. (2011). A realização da técnica prosseguiu a partir do uso da enzima *EcoT22I*, escolhida pois apresentou o melhor perfil de digestão do DNA genômico de *H. brasiliensis*. Para o sequenciamento das bibliotecas, o sequenciador Illumina GAIIx (Illumina, San Diego, CA, USA) foi utilizado.

Análise dos dados foi realizada através do TASSEL-GBS 3.0 e 4.0 pipeline (GLAUBITZ et al., 2014). As sequências do genoma de *Hevea brasiliensis* (TANG et al., 2016) foram usadas como referência para alinhamento por meio do Bowtie2 versão 2.1 (LANGMEAD; SALZBERG, 2012). Para os *reads* obtidos previamente, foi estabelecido que uma *tag* deveria estar repetida no mínimo seis vezes (*default* = 1). Ao fim do processo, todas as variantes foram exportadas como arquivos VCF e filtradas para o máximo de 25% de dados perdidos por marcador e presença apenas de SNPs bialélicos através do *software* vcftools (DANECEK et al., 2011).

Mapeamento Genético

A construção do mapa genético foi desenvolvida em duas etapas. Na primeira, marcadores microssatélites e SNPs (com exceção aos de GBS) foram mapeados utilizando LOD de 5,37 e frequência de recombinação máxima de 0,4 por meio do *software* OneMap versão 2.0-4 (MARGARIDO; SOUZA; GARCIA, 2007), implementado como um pacote adicional ao software estatístico R (R CORE TEAM, 2015). Em seguida, o *software* OneMap versão 2.0-6 (em desenvolvimento), que permite analisar marcadores baseados nas tecnologias de NGS, foi utilizado para filtragem final e inserção dos SNPs provenientes do método de GBS no mapa previamente construído.

Análises de SNPs obtidos por meio da abordagem de GBS foram realizadas para identificar marcadores duplicados, marcadores com informação redundante e marcadores com desvio na segregação esperada. Utilizando LOD de 7,22 e fração de recombinação de 0,4 os marcadores foram mapeados, sendo que o teste para a ordenação dos SNPs foi realizado tomando por base a inserção de cada marcador na estimativa de sua melhor posição. Para a determinação das distâncias em centiMorgans foi utilizado a função de mapeamento de Kosambi (KOSAMBI, 1943). Por fim, os grupos de ligação do mapa final foram visualizados usando o *software* Mapchart 2.1 (VOORRIPS, 2002).

Análise Fenotípica

O diâmetro do caule foi medido a 50 cm do calo de enxertia durante as estações de inverno e verão. As análises ocorreram em cinco pontos de tempo com intervalos de aproximadamente 6 meses, sendo elas em julho de 2013 (12 meses), dezembro de 2013 (17 meses), maio de 2014 (22 meses), novembro de 2014 (28 meses) e junho de 2015 (35 meses). A análise de modelos em um primeiro momento, consistiu no ajuste do modelo apropriado com a obtenção dos componentes de variância pela maximização da distribuição de máxima verossimilhança restrita (REML). A decomposição dos valores fenotípicos em efeitos principais genéticos e de delineamentos podem ser constatados no seguinte modelo abaixo:

$$y = X_T t + Z_R r + Z_B b + Z_G g + \varepsilon$$

Em que y é referente aos dados fenotípicos avaliados (escala em cm); X_T é a matriz de incidência do efeito sistemático de tempo (intercepto do modelo multivariado - efeito fixo); t é vetor de efeitos fixos de época; Z são as matrizes de incidência dos efeitos aleatórios; r é o vetor de efeitos aleatórios de repetição aninhados dentro de época $r \sim MVN(0, I_R \otimes G_R)$; b representando o vetor de efeitos aleatórios de blocos aninhados dentro de repetição aninhados em época $b \sim MVN(0, I_B \otimes I_R \otimes G_B)$; g é o vetor de efeitos aleatórios do genótipo aninhados dentro de época $g \sim MVN(0, A \otimes G_G)$; ε é o vetor de resíduos $\varepsilon \sim MVN(0, R)$; A é a matriz de parentesco baseado em genealogia (numerator relationship matrix (MRODE, 2014); I_R e I_B são as matrizes de identidade de ordem 4 para o número de repetições e número de blocos. No modelo testou-se diferentes estruturas de variância-covariância para as matrizes genéticas G_R , G_B , G_G e também para a matriz de variância-covariância do resíduo (R), sendo esta, construída a partir do modelo de linhas e colunas visando controlar a variabilidade espacial do experimento (SMITH; CULLIS; THOMPSON, 2005).

As análises foram realizadas utilizando os pacotes ASReml-R (BUTLER et al., 2009) e ASRemlPlus (BRIEN, 2016) no ambiente R. Valores genéticos preditos (BLUPs) foram usados para a realização do mapeamento de QTLs. O cálculo da herdabilidade no sentido restrito (h^2) foi baseado nas estimativas dos componentes de variância, onde: $h^2 = \hat{\sigma}_{A(t)}^2 / \sigma_{(t)}^2$, sendo $\hat{\sigma}_{A(t)}^2$, a variância aditiva no tempo t estimada pelo modelo, e $\sigma_{(t)}^2$ a variância fenotípica total no tempo t .

Mapeamento de QTLs

Análises de QTLs foram realizadas utilizando o método de mapeamento por intervalo composto (ZENG, 1994), conforme modelo descrito por Gazaffi et al. (2014) para população F1 segregante, com probabilidade condicional multiponto para o genótipo dos QTLs obtidas usando o *software* OneMap versão 2.0-4 (MARGARIDO; SOUZA; GARCIA, 2007).

Para a realização do mapeamento de QTLs, se fez necessário estabelecer uma hipótese que um QTL está localizado dentro de um intervalo em consideração, com marcadores (fora deste intervalo) ligados ao QTL putativo, incluídos como cofatores para assim, poderem contribuir na acurácia do mapeamento. Dessa forma, foi realizado o teste ao longo do genoma em intervalos regulares de 1 cM, se evitando uma janela de 15 cM ao redor da posição sob avaliação. A consideração sobre a

presença de um QTL foi analisada por meio de valores de LOD score, com nível de significância de 95% para 1000 permutações conforme proposto por Churchill e Doerge (1994) e modificado por Chen e Storey (2006). Ainda, para cada modelo CIM, foram estimados os efeitos genéticos aditivos para cada genitor e efeitos de dominância, bem como fases de ligação, padrões de segregação e a proporção da variação fenotípica explicada pelos QTLs (R^2). Todas as análises foram realizadas no ambiente R (R CORE TEAM, 2015) usando um pacote em desenvolvimento.

Resultados e Discussão

População de mapeamento

A obtenção de uma população de mapeamento gerada a partir de polinização aberta evitou dificuldades devido à baixa taxa de sucesso para polinização controlada na espécie (PRIYADARSHAN; CLÉMENT-DEMANGE, 2004). Esta abordagem foi possível, devido a presença de uma ampla área experimental contendo clones GT1 e RRIM701 lado a lado, e também por meio das características biológicas do clone GT1, que devido a rearranjos estruturais em sequências envolvidas com a síntese de ATP, tem energia insuficiente para o desenvolvimento de anteras, caracterizando-se como macho-estéril e se comportando apenas como árvore mãe (SHEARMAN et al., 2014).

O uso de uma progênie oriunda de polinização aberta a partir de uma árvore mãe de interesse também foi descrito por Guajardo et al. (2015) em cereja doce, e busca facilitar o desenvolvimento de uma progênie e assim, acelerar a construção de mapas e localização de QTLs. A genotipagem dos 526 marcadores confirmou que a identificação do doador de pólen para formação da população de mapeamento foi realizada com sucesso, pois, com exceção de dois indivíduos, os alelos observados na progênie correspondem a segregação esperada a partir do cruzamento entre GT1 e RRIM701.

Filtragem dos marcadores e análise do polimorfismo

A partir de 1.174 marcadores SSRs avaliados para o polimorfismo entre os genitores da população de mapeamento, foi possível observar 569 marcadores polimórficos (~48%), dos quais 281 foram EST-SSRs. Os demais marcadores foram considerados monomórficos ou não puderam ser analisados devido a falha na amplificação ou ausência de bandas nítidas. Níveis de polimorfismo similares foram detectados em outras populações de mapeamento de *Hevea brasiliensis* por Souza

et al. (2013) e Lespinasse et al. (2000). Dentre os marcadores EST-SSR, maior proporção de polimorfismo para a população foi detectada em comparação com os marcadores avaliados por Nirapathpongporn et al. (2016).

Por sua vez, os 487 marcadores SNPs selecionados para genotipagem por meio de diferentes metodologias, mostraram polimorfismo entre os genitores da população de mapeamento em aproximadamente 49% das ocasiões (239 SNPs). Tais plataformas de genotipagem têm sido usadas em outras espécies com genoma complexo, como cana de açúcar (COSTA et al., 2016) e algodão (BYERS et al., 2012), pois têm possibilitado não só o processamento de grande quantidade de dados de maneira rápida, mas também, detectar e quantificar com precisão alelos com frequências muito baixas (OBERACHER, 2008; WANG et al., 2009).

O teste de Chi-quadrado realizado nos 526 marcadores genotipados na população de mapeamento, revelou que 156 marcadores (29,6%) possuem segregação 1:1:1:1, 133 (25,3%) segregação 1:2:1 e 237 (45,1%) segregação 1:1. A seleção de 289 marcadores com maior poder informativo (segregação 1:1:1:1 e 1:2:1) é de grande importância nas análises de ligação devido a limitação no uso de marcadores com segregação 1:1, que fornecem informação insuficiente para distinguir genótipos recombinantes e não-recombinantes (WU et al., 2002).

Todos os marcadores genotipados foram mantidos nas análises de ligação, incluindo 33 marcadores (6,3%) identificados com distorção da razão de segregação mendeliana esperada ($P \leq 0,005$). Isso foi possível, uma vez que este fenômeno é comum em análises genéticas (XU, 2008) e apresenta efeito pequeno sobre populações durante a construção de mapas de ligação (HACKETT; BROADFOOT, 2003). A presença dessa classe de marcadores não levou a alterações na formação dos grupos de ligação do presente trabalho.

A genotipagem por sequenciamento gerou um total de 68.777.856 *reads* brutos obtidos a partir das bibliotecas *paired-end*, com uma média de 474.330 *reads* para cada uma das amostras. Após o alinhamento com o genoma de referência de *Hevea brasiliensis* (TANG et al., 2016) utilizando uma profundidade mínima de seis tags, um total de 81.832 SNPs foram identificados. Posteriormente, a seleção de marcadores com no máximo 25% de dados perdidos resultou em 28.670 SNPs. Após as filtragens

realizadas por meio do *software* Onemap versão 2.0-6 foram identificados 2.998 marcadores de alta qualidade.

Esforços na filtragem de dados de GBS para obtenção de marcadores robustos têm sido realizados para diversas espécies (COVARRUBIAS-PAZARAN et al., 2016; GUAJARDO et al., 2015; LU et al., 2013; WARD et al., 2013). Em seringueira, Pootakham et al. (2015) utilizando uma profundidade mínima de tags ≥ 6 e menos de 50% de dados perdidos obtiveram em dois cruzamentos analisados, 7.345 e 6.678 SNPs e, posteriormente, após utilizar parâmetros mais restritivos (profundidade mínima de tags ≥ 30 e 10% de dados perdidos) e separar marcadores polimórficos em pelo menos um dos genitores de cada população, um total de 2.995 e 3.124 SNPs. Tais quantidades são similares aos obtidos no presente trabalho e demonstram a eficiência dos parâmetros utilizados na seleção dos SNPs.

Dentre os marcadores selecionados, 627 SNPs (20,9%) foram classificados com segregação 1:2:1 e 2.371 SNPs (79,1%) classificados com segregação 1:1. Shearman et al. (2015), encontraram dificuldades na identificação de SNPs devido à elevada heterozigosidade natural de *Hevea brasiliensis* após sequenciamento de exoma, com porcentagens de genótipos heterozigotos nos genitores (63 e 71%) similares ao presente trabalho. Comparações do número de locos heterozigotos nos genitores obtidos Pootakham et al. (2015), são difíceis, uma vez que os valores encontrados variando entre 32,3 a 40,8%, não consideram análises de distorção de segregação.

A complexa identificação de SNPs em espécies que sofreram eventos de poliploidização, como possivelmente ocorreu em *Hevea brasiliensis*, torna a escolha de estratégias adequadas para seleção de SNPs úteis e informativos ao mapeamento um processo necessário (CLEVENGER et al., 2015). A redução no número final de SNPs obtidos (2.998 SNPs) reflete o uso de parâmetros restritivos utilizados na seleção dos marcadores, e é importante, pois os erros de genotipagem e a grande quantidade de dados perdidos gerados pela técnica (DAVEY et al., 2011), têm efeitos na ordenação e estimativa da distância entre locos (HACKETT; BROADFOOT, 2003; SHIELDS et al., 1991) e, conseqüentemente, podem afetar a localização de regiões que controlam caracteres quantitativos (DOERGE; ZENG; WEIR, 1997).

Mapeamento Genético

Análises de ligação iniciais utilizando 526 marcadores (287 SSRs e 239 SNPs), LOD score de 5,37 e fração de recombinação de 0,4 identificaram 18 grupos de ligação, com um total de 411 marcadores distribuídos em 2.482,3 cM (Figura 3). O comprimento de cada grupo variou de 78,3 cM para o grupo de ligação 4 a 191,9 cM para o grupo de ligação 14. Por sua vez, o grupo 10 apresentou a maior cobertura e o maior número de marcadores mapeados, enquanto o grupo de ligação 16 a menor densidade (Tabela 1).

A organização dos grupos de ligação foi baseada nos mapas de ligação construídos anteriormente para *Hevea brasiliensis* (LESPINASSE et al., 2000; SOUZA et al., 2013), possibilitando realizar comparações sobre o ordenamento de marcadores mapeados em comum. Com comprimento de 228,7 cM e 20 marcadores mapeados, o maior grupo de ligação obtido por Souza et al. (2013) corresponde ao grupo 10, apresentando um total de 10 marcadores em comum, sendo dois (HB135 e HB174) em posições diferentes. Embora não tenha sido o maior grupo de ligação descrito por Souza et al. (2013), o grupo de ligação 14 apresentou o mesmo ordenamento entre nove marcadores mapeados em comum, com um total de 217,4 cM.

Tabela 1: Informações sobre os 18 grupos de ligação da população GT1 x RRIM701.

LG*	Mapa Base					Mapa Final				
	SSRs	SNPs	Compr. (cM)	Densid. (cM)	Gap (cM)	SSRs	SNPs	Compr. (cM)	Densid. (cM)	Gap (cM)
1	14	17	153,1	4,9	18,4	14	53	232,4	3,5	12,9
2	19	8	140,2	5,2	14,4	19	33	199,3	3,8	9,1
3	14	6	115,0	5,8	20,7	14	43	184,9	3,2	14,0
4	6	5	78,3	7,1	17,9	6	39	163,8	3,6	12,9
5	9	13	176,9	8,0	24,9	9	65	272,0	3,7	16,6
6	17	14	162,9	5,3	20,1	17	63	228,1	2,9	11,6
7	12	5	80,3	4,7	19,4	12	32	109,4	2,5	14,1
8	14	8	141,3	6,4	18,4	14	45	229,7	3,9	18,4
9	13	6	124,6	6,6	20,8	13	33	196,3	4,3	14,3
10	26	27	185,4	3,5	16,2	26	52	228,1	2,9	11,5
11	7	14	164,8	7,9	27,5	7	57	235,4	3,7	20,6
12	7	10	104,1	6,1	23,3	7	32	167,5	4,3	12,8
13	6	10	136,7	8,5	24,2	6	58	218,7	3,4	15,0
14	22	13	191,9	5,5	15,6	22	51	258,8	3,6	17,0
15	10	6	160,0	10,0	32,5	10	39	225,3	4,6	16,2
16	10	6	100,8	6,3	17,3	10	53	166,0	2,6	12,9
17	6	7	107,5	8,3	19,6	6	54	206,7	3,5	9,8
18	14	10	158,5	6,6	26,0	14	54	257,3	3,8	17,1
Total	226	185	2482,3	6,0	32,5	226	856	3779,7	3,5	20,6

LG* – Grupo de ligação; Compr. – Comprimento Médio em cM; Densid. – Densidade de marcadores em cM; Gap – Comprimento do maior gap em cM

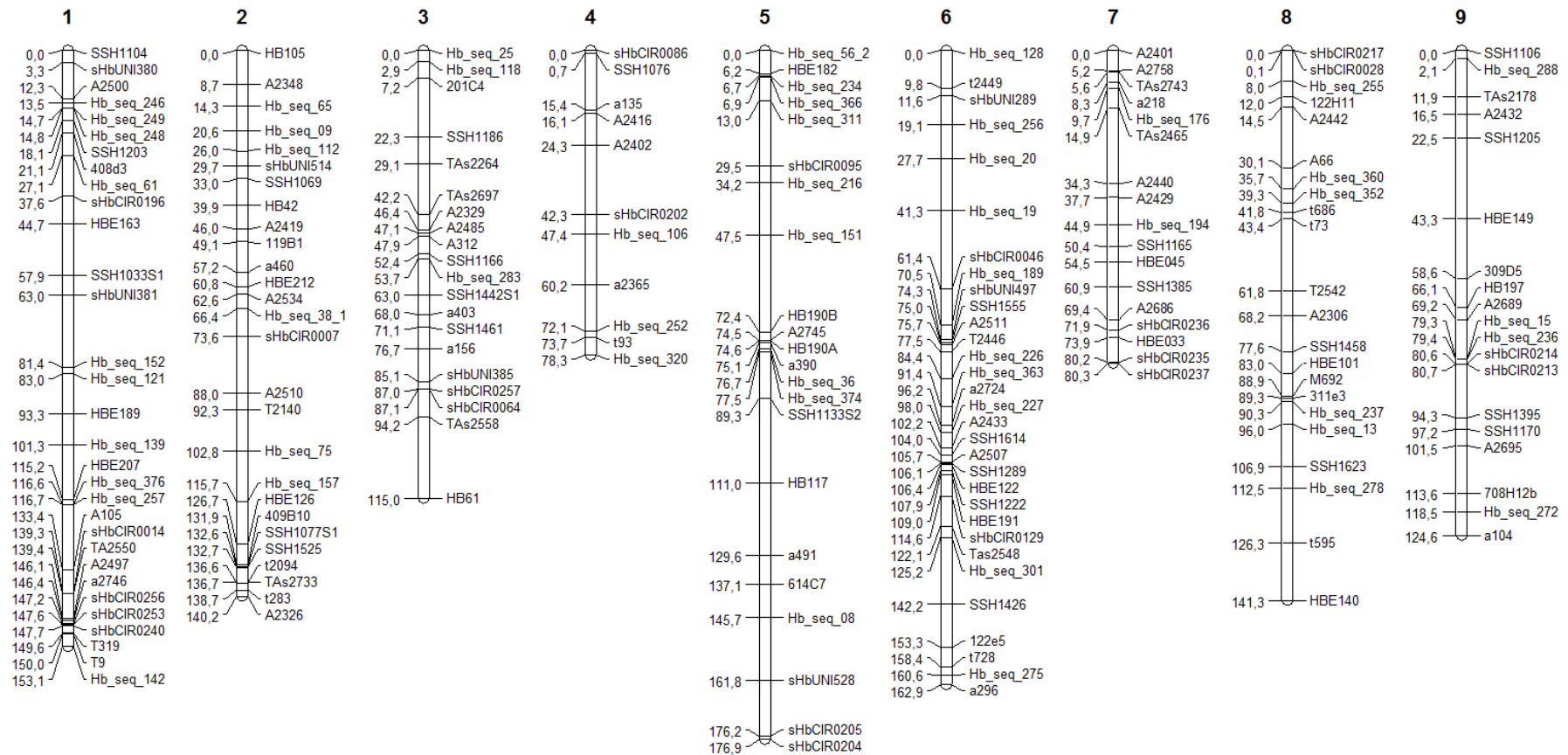


Figura 3: Grupos de ligação 1 a 9 do mapa genético base contendo 411 marcadores (226 SSRs e 185 SNPs).

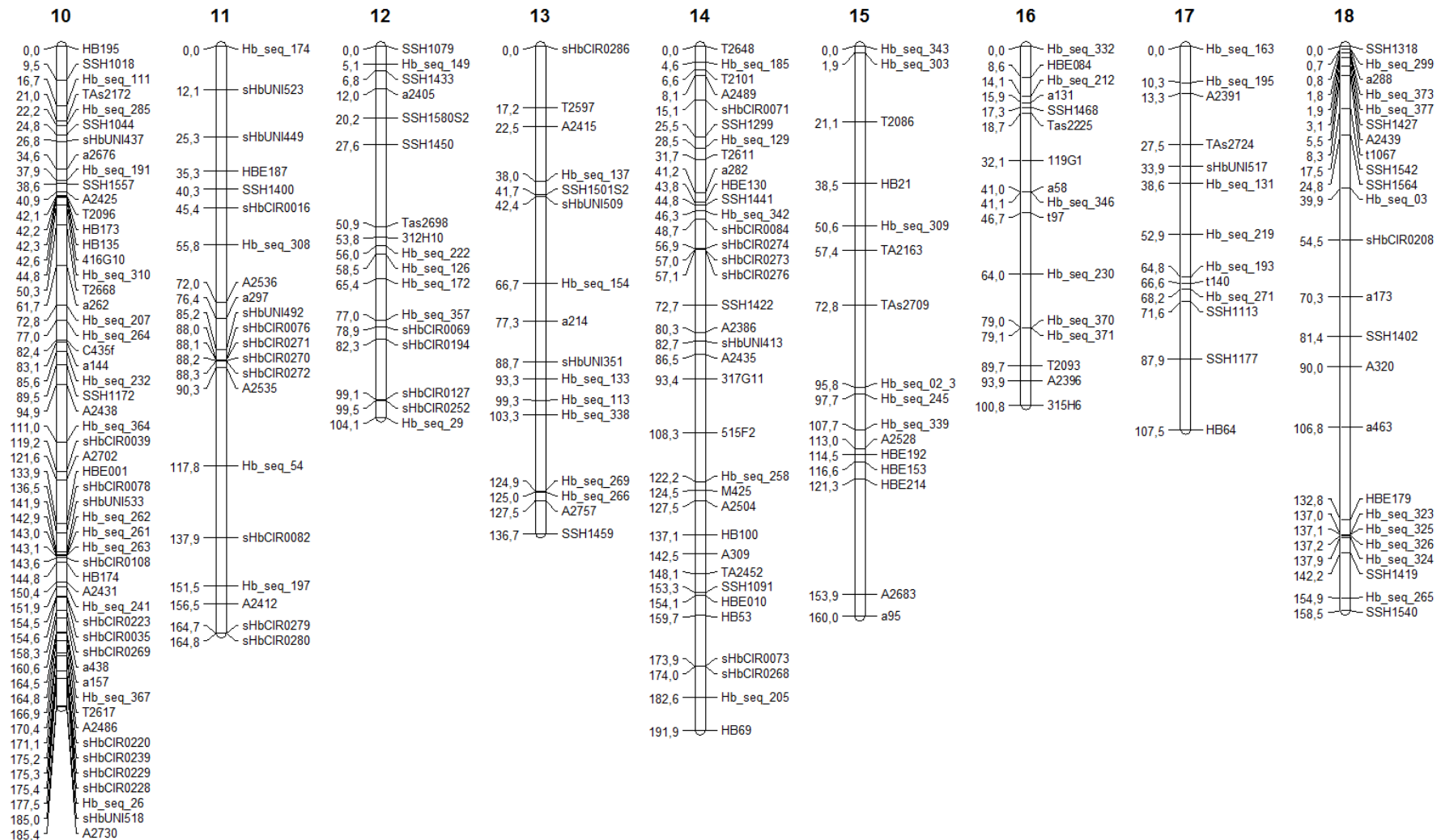


Figura 3 (continuação): Grupos de ligação 10 a 18 do mapa genético base contendo 411 marcadores (226 SSRs e 185 SNPs)

Assim como no presente trabalho, o mapeamento com diferentes tipos e quantidades de marcadores tem apontado para a ocorrência de 18 grupos de ligação em seringueira (LE GUEN et al., 2011a; LESPINASSE et al., 2000; NIRAPATHPONGPORN et al., 2016; POOTAKHAM et al., 2015; SHEARMAN et al., 2015), correspondendo ao número haploide para a espécie ($2n = 36$). Os grupos de ligação obtidos com 226 SSRs e 185 SNPs corroboram as observações feitas por Souza et al. (2013) quanto a necessidade de utilizar diferentes marcadores para que grupos de ligação adicionais, oriundos da cobertura incompleta do genoma de *Hevea brasiliensis*, possam ser resolvidos.

Apesar da densidade média de marcadores no mapa genético prévio ter sido maior que o observado por Triwitayakorn et al. (2011) e Souza et al. (2013) ao mapearem 97 e 284 SSRs respectivamente, intervalos entre marcadores superiores a 15 cM puderam ser observados em 41 ocasiões. Assim, da mesma forma como levantado por Souza et al. (2013), tais resultados podem ser explicados pela baixa frequência de recombinação em regiões com elevado grau de homozigose ou pelo menor número de marcadores polimórficos nessas regiões, reforçando a necessidade de aumento da resolução por meio de outras metodologias.

Por meio de grande quantidade de SNPs de alta qualidade facilmente produzidos com a técnica de GBS, o mapa genético final foi construído utilizando o mapa genético base obtido anteriormente. Assim, um total de 1.082 marcadores distribuídos ao longo de 18 grupos de ligação cobriram um total de 3.779,7 cM (Figura 4). Em média, cada grupo conteve aproximadamente 60 marcadores cobrindo cerca de 210 cM. O maior grupo de ligação (LG5) incluiu 74 marcadores abrangendo 272 cM com uma distância média entre marcadores de 3,68 cM, e o menor (LG7), conteve 44 marcadores cobrindo 109,4 cM, representando também o grupo com a menor distância média entre marcadores (2,49 cM).

A utilização de um mapa genético base constituído de marcadores SSRs, foi essencial para evitar possíveis problemas na formação de grupos de ligação, ordenamento e cálculo das estimativas da distância de marcadores. Também, a possibilidade de inserir grande quantidade de marcadores SNPs para contornar problemas de regiões altamente homozigóticas e com baixa densidade de marcadores, como observado no mapa base e destacado por Souza et al. (2013), demonstrou ser possível, pois reduções consideráveis no intervalo médio entre marcadores foi observado para todos

os grupos de ligação, com média de 3,49cM (Tabela 1). Aumento da resolução nessas regiões foram encontrados para a espécie em outras oportunidades (POOTAKHAM et al., 2015; SHEARMAN et al., 2015), entretanto sem considerar os efeitos benéficos dos marcadores SSRs, tal como realizado em uma espécie do gênero *Brassica* (YANG et al., 2016).

A observação da distribuição de marcadores SNPs de GBS ao mapa genético (marcadores em preto na figura 4), demonstra que o mapeamento foi bem conduzido, uma vez que as marcas são encontradas uniformemente ao longo dos grupos de ligação. Não se nota no mapa, grandes blocos de marcadores em regiões de espaço limitado, mas sim o preenchimento de *gaps* antes observados (Figura 3). Nesse sentido, se destaca a redução no número total de *gaps* acima de 15 cM, antes verificados em 41 ocasiões, para apenas oito, reforçando a eficiência da abordagem de GBS no mapeamento de *Hevea brasiliensis*.

Regiões de grupos de ligação com baixa densidade de marcadores, como por exemplo, nos grupos de ligação 13 e 15, que possuíam *gaps* de até 24,2 cM e 32,5 cM no mapa prévio, puderam ser mais saturadas. Por outro lado, assim como em outros trabalhos de mapeamento genético utilizando metodologias de genotipagem em larga escala para *Hevea brasiliensis* (POOTAKHAM et al., 2015; SHEARMAN et al., 2015), intervalos consideráveis (≥ 10 cM) continuaram a ser observados, corroborando com os resultados obtidos com plataformas similares no mapeamento de diferentes espécies vegetais, como *Vigna vexillata* (MARUBODEE et al., 2015), mirtilo (MCCALLUM et al., 2016), ervilha (BOUTET et al., 2016) e cereja doce (GUAJARDO et al., 2015).

A presença de *gaps* reforça a complexidade do processo de mapeamento genético, especialmente para espécies perenes, pois o aumento do número de eventos de recombinação através da obtenção de populações muito grandes é inviável (FIERST, 2015). Fica claro também que a interação entre genética e genômica, bem como metodologias genético-estatísticas devem ser utilizadas para detecção e localização QTLs (DOERGE; ZENG; WEIR, 1997).

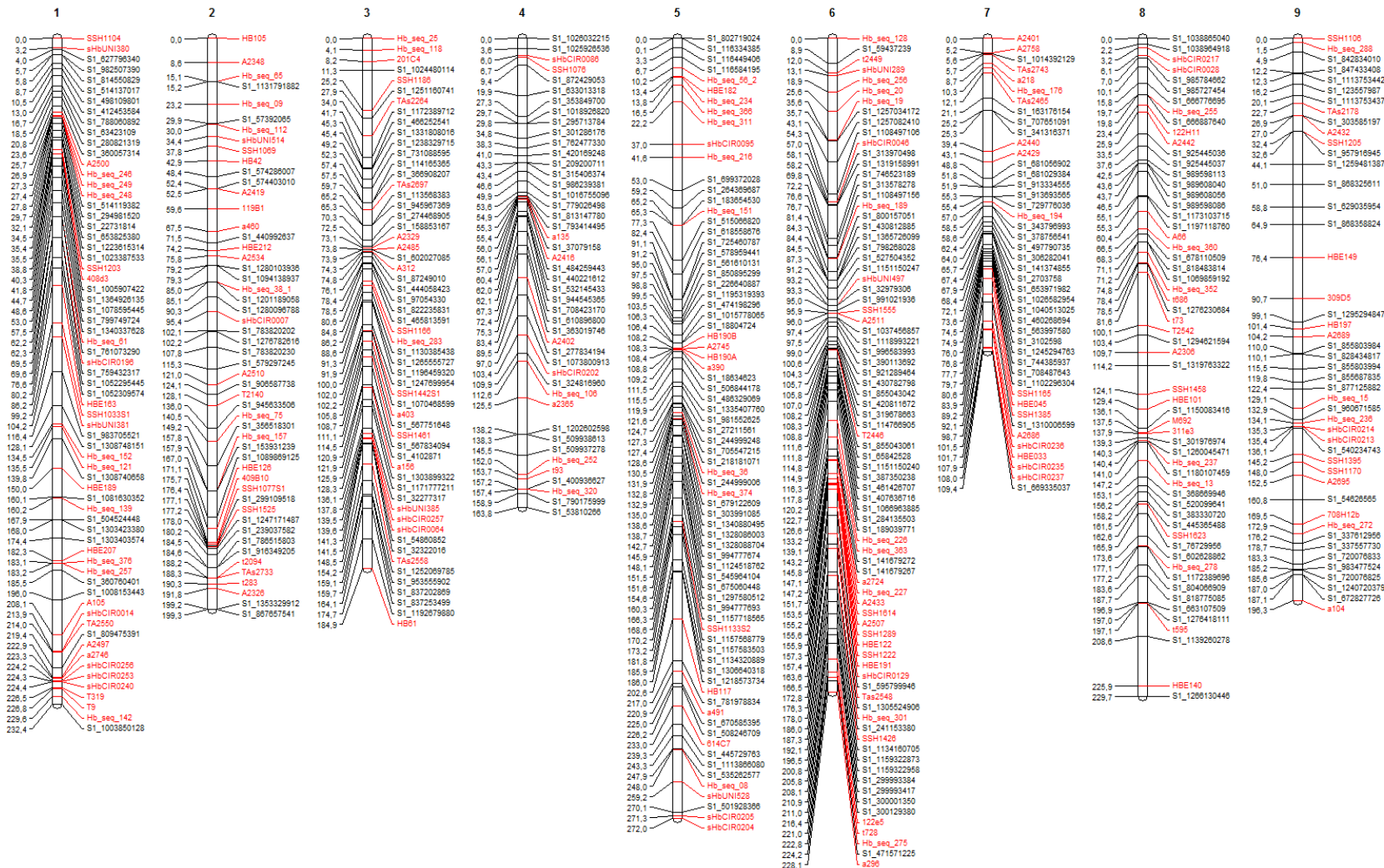


Figura 4: Grupos de ligação 1 a 9 do mapa genético final para população GT1 x RRM701. Em preto, marcadores SNPs de GBS e em vermelho marcadores mapeados no mapa base.

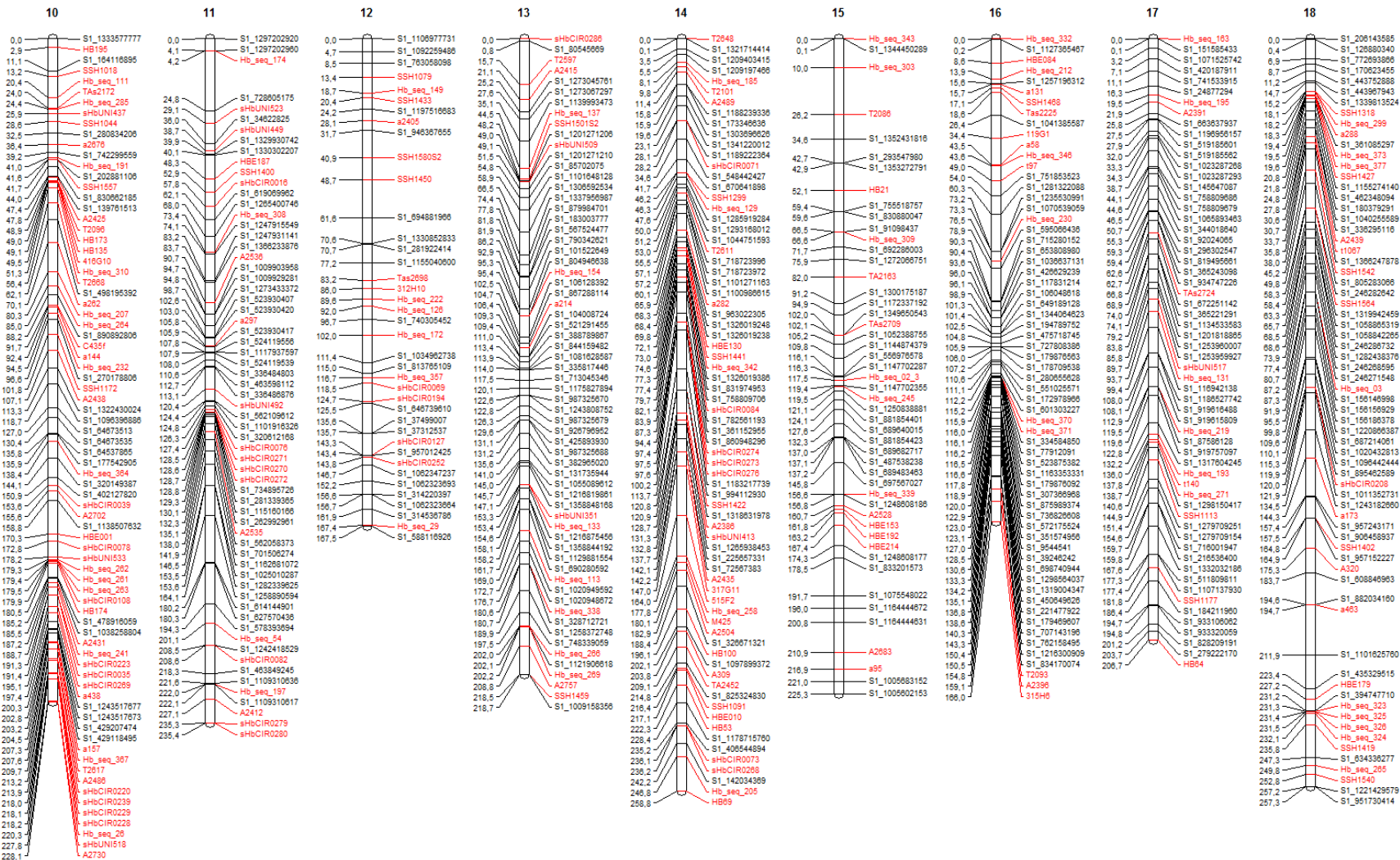


Figura 4 (continuação): Grupos de ligação 10 a 18 do mapa genético final para população GT1 x RRIM701. Em preto, marcadores SNPs de GBS e em vermelho marcadores mapeados no mapa base.

Análise Fenotípica

A análise dos dados fenotípicos realizada por meio da abordagem de modelos mistos incluiu a comparação de diferentes estruturas de variância-covariância dos efeitos aleatórios (Tabela 2). Tomando por base os menores valores do Critério de Informação de Akaike (AIC) (AKAIKE, 1974) e do Critério de Informação Bayesiana (BIC) (SCHWARZ, 1978), o modelo escolhido (em negrito na tabela 2) considerou uma estrutura de variância-covariância heterogênea para os efeitos genéticos aleatórios.

Tabela 2: Resumo dos modelos testados para o experimento.

Modelo	G_R	G_B	G_G	R_C	R_L	R_T	gl	AIC	BIC
1	Fixo	Fixo	ID	ID	ID	ID	2	37234,1	37248,0
2	Fixo	Fixo	ID	ID	ID	ID	2	37299,1	37313,0
3	ID	ID	ID	ID	ID	ID	4	37527,7	37555,4
4	Ar1v	Ar1v	Ar1v	ID	ID	ID	7	36900,1	36948,6
5	Corh	Corh	Corh	ID	ID	ID	19	36720,0	36851,8
6*	Ar1h	Ar1h	Ar1h	ID	ID	ID	-	-	-
7*	US	US	US	ID	ID	ID	-	-	-
8	Corh	Corh	Corh	Diag	ID	ID	23	35140,7	35300,2
9	Corh	Corh	Corh	ID	ID	Corh	24	29619,5	29785,9
10	Corh	Corh	Corh	ID	ID	Ar1h	24	29413,5	29579,8
11	Corh	Corh	Corh	Ar1	Ar1	Ar1h	26	29329,9	29510,1
12	Ar1v	Ar1v	Corh	Ar1	Ar1	Ar1h	18	29351,5	29476,3
13	Ar1v	Ar1v	Corh	Ar1	Ar1	Corh	18	29564,5	29689,3
14	Corh	Corh	Corh	Ar1	Ar1	Diag	25	35032,4	35205,7
15	Diag	Diag	Corh	Ar1	Ar1	Diag	23	35098,7	35258,2
16	Diag	Diag	Ar1h	Ar1	Ar1	Diag	22	35659,1	35911,6
17	Ar1v	Ar1v	Corh	Ar1	Ar1	Diag	17	35049,3	35167,2
18	Ar1v	Ar1v	Ar1h	Ar1	Ar1	Corh	18	29560,4	29685,2

$G_{(R,B,G)}$: matriz variância-covariância para efeitos de repetição, blocos e genótipos

$R_{(C,L,T)}$: matriz variância-covariância do resíduo para efeitos de coluna, linha e tempo

gl: graus de liberdade

Fixo: se o efeito foi considerado fixo

ID: matriz identidade (variância genética homogênea sem correlação)

Diag: matriz diagonal (variância genética heterogênea sem correlação)

Ar1v: matriz de correlação autoregressiva de 1ª ordem com variância homogênea

Corh: matriz de variância heterogênea com correlação

US: matriz de variância-covariância não-estruturada

Ar1h: matriz de correlação autoregressiva de 1ª ordem com variância heterogênea

Ar1: matriz de correlação autoregressiva de 1ª ordem sem variância própria

* Modelos que não convergiram devido a singularidade das matrizes e/ou muitos parâmetros estimados

Considerando as estruturas do modelo em foco, é possível afirmar que cada período de tempo possui uma variância única e que a covariância entre os tempos possui uma correlação. Por sua vez, em relação às matrizes de variância-covariância

genética dos efeitos de repetição e blocos, espera-se que em cada diferente período de tempo, e de acordo com variações de temperatura e quantidade de chuvas, o ambiente do experimento como um todo responda de uma diferente forma, proporcionando assim, variações únicas, porém correlacionadas com o tempo anterior. Tais estruturas de variância-covariância heterogênea com correlação pode ser considerada adequada para modelar interações genótipo-ambiente (PAGET et al., 2015).

Em relação as estruturas das matrizes de resíduo para os efeitos de tempo, a mesma foi ajustada com estrutura que supõe cada tempo com uma variância distinta (da mesma forma que efeitos aleatórios) e que a correlação entre elas diminui com o tempo. Assim, efeitos não explicados pelo modelo (efeitos ambientais, de genótipo não-aditivos, de erros de medição) são correlacionados ao longo do tempo, porém esta correlação tende a diminuir. Já as estruturas das matrizes de resíduo para os efeitos de coluna e linha, foram ajustadas com estrutura de correlação, que supõe que um dado resíduo de determinada árvore possui influência em sua vizinhança e essa influência diminui com a distância entre as árvores, que pode ser causado por variações ambientais como, por exemplo, pequenas manchas de solo e/ou de adubação, erros sistemáticos de medição e diferenças de poucos dias na data de implantação do experimento.

Assim, a partir do modelo ajustado, se prosseguiu para a obtenção dos valores genéticos preditos a partir da metodologia conhecida como BLUP (*Best Linear Unbiased Prediction*) (HENDERSON, 1973) e acurácia da predição dos valores genéticos (COWLING et al., 2015; MRODE, 2014). Descrições dos valores de acurácia da predição dos valores genéticos são observados na tabela 3, de forma que tais valores representam a correlação entre os valores preditos e o que pode ser estimado (COWLING et al., 2015).

Tabela 3: Acurácia da predição dos valores genéticos nos diferentes tempos.

	12 meses	17 meses	22 meses	28 meses	35 meses
Mínimo	0,56	0,61	0,59	0,65	0,67
1º Quartil	0,77	0,81	0,80	0,83	0,86
Mediana	0,78	0,82	0,81	0,83	0,87
Média	0,77	0,82	0,80	0,83	0,86
3º Quartil	0,78	0,82	0,81	0,84	0,87
Máximo	0,88	0,92	0,90	0,93	0,96

A análise da herdabilidade no sentido restrito (h^2) revelou maiores valores para os períodos de 12, 17 e 22 meses (Tabela 4), enquanto que valores menores foram observados para 28 meses e 35 meses. Tais resultados demonstram que parte da variação fenotípica também deve ser atribuída a diferenças nos genótipos da progênie. A partir dos valores fenotípicos observados nos genitores (GT1 e RRIM701) se constatou que nenhum se sobressaiu durante os períodos avaliados para o diâmetro, enquanto que valores maiores puderam ser encontrados na progênie.

Tabela 4: Resumo dos valores genotípicos preditos, estimativa dos componentes de variância genética aditiva ($\hat{\sigma}_A^2$) e fenotípica ($\hat{\sigma}_P^2$), coeficiente de variação genotípica (CV_G) e herdabilidade no sentido restrito (h^2).

Característica/Tempo	Valores Genotípicos Preditos					$\hat{\sigma}_A^2$	$\hat{\sigma}_P^2$	CV_G	h^2
	Mínimo	Máximo	Média	GT1	RRIM701				
Diâmetro/12 meses	13,88	19,94	17,05	17,24	16,89	3,21	20,79	0,06	0,15
Diâmetro/17 meses	17,77	25,23	21,69	21,79	21,66	5,19	31,81	0,06	0,16
Diâmetro/22 meses	24,09	34,15	29,35	29,77	29,16	8,62	55,01	0,06	0,16
Diâmetro/28 meses	32,08	38,4	35,47	35,86	35,03	3,87	74,41	0,03	0,05
Diâmetro/35 meses	43,56	58,08	51,17	52,01	50,44	18,20	148,01	0,05	0,12

Mapeamento de QTLs

A análise de QTLs foi realizada com base nos valores genéticos preditos, através do método de mapeamento por intervalo composto (CIM) para progênie de cruzamentos F1 segregantes (GAZAFFI et al., 2014). As análises foram conduzidas separadamente para cada tomada de dados (12, 17, 22, 28, 35 meses) (Figura 5 e Tabela 5).

Foram identificados um total de 19 QTLs ao longo dos estágios de desenvolvimento (Tabela 5). Utilizando um LOD score de 4,18 o maior número de QTLs foi identificado na avaliação aos 12 meses (seis QTLs detectados), que corresponderam juntos por 47,05% da variação fenotípica. Por outro lado, dois QTLs foram observados aos 28 meses na avaliação de novembro de 2014 nos grupos de ligação 2 e 17, explicando um total de 16,94% da variação fenotípica. Ao todo, foi possível identificar QTLs nos grupos de ligação 2, 5, 6, 8, 14, 17 e 18, e a nomeação dos mesmos, seguiu o prefixo da característica avaliada com os tempos das mensurações fenotípicas, seguido do grupo de ligação; por exemplo, Diam12-GL6.

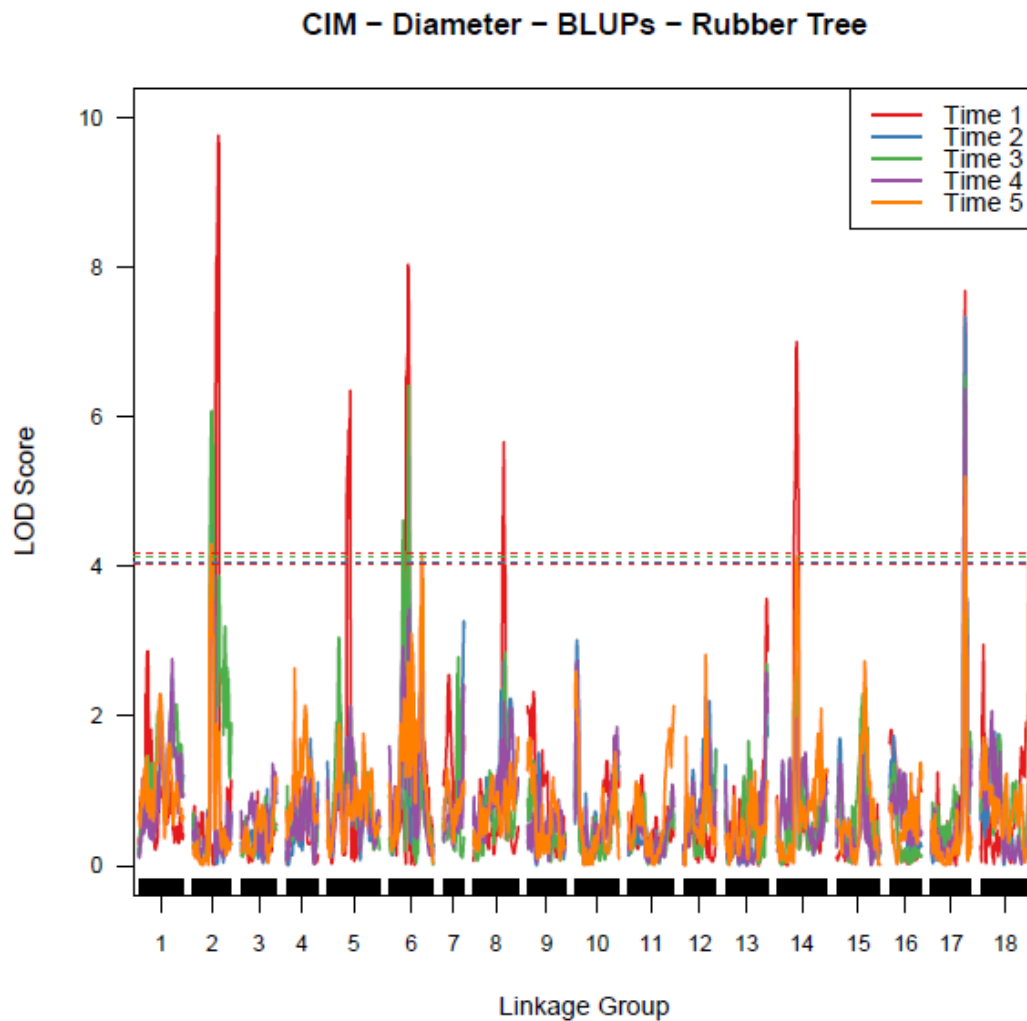


Figura 5: Mapeamento de QTLs para diâmetro em diferentes tempos avaliados. Valores de *threshold* para cada tempo obtidos com 1000 permutações.

Uma comparação entre as diferentes estações do ano também foi realizada. Os QTLs detectados durante o verão foram diferentes daqueles detectados durante o inverno. Durante o inverno encontramos dois QTLs (posições 95,4cM no grupo de ligação 2 e 176,0 cM no grupo de ligação 17) ocorrendo em ambos os tempos avaliados (17 e 28 meses). O QTL identificado no grupo de ligação 2 apresentou segregação 3:1 com efeito aditivo em ambos genitores e efeito de dominância; o possível QTL no grupo de ligação 17 apresentou segregação 1:2:1, efeito significativo para dominância e aditividade com relação ao genitor GT1.

Tabela 5: Resultados do mapeamento de QTLs para diâmetro no cruzamento GT1 x RRM701. Avaliações fenotípicas mensuradas em 12, 17, 22, 28 e 35 meses.

QTL	LG	Posição (cM)	LOD	R ²	Efeito Aditivo GT1	LOD ⁽¹⁾	Efeito Aditivo RRM701	LOD ⁽²⁾	Efeito de Dominância	LOD ⁽³⁾	Segregação
Diam12-2	2	132,00	9,77	13,68	0,24	3,25	0,03	0,03	0,41	6,35	1:2:1
Diam12-5	5	118,00	6,36	6,23	-0,02	0,03	-0,35	5,08	-0,14	0,84	1:2:1
Diam12-6	6	97,40	8,04	12,09	-0,27	3,67	-0,04	0,09	0,32	5,28	1:2:1
Diam12-8	8	156,20	5,66	5,40	-0,30	4,48	0,04	0,11	0,13	0,90	1:2:1
Diam12-14	14	100,22	7,00	10,95	0,14	0,89	0,00	0,00	-0,37	6,37	1:1:1:1
Diam12-17	17	176,00	7,69	11,26	0,28	3,51	0,13	1,03	-0,34	4,98	1:1:1:1
Diam17-2	2	95,40	4,99	8,09	0,37	3,06	-0,26	1,75	0,27	1,65	3:1
Diam17-6	6	100,55	5,47	7,65	-0,78	4,52	-0,35	1,14	0,10	0,15	1:1:1:1
Diam17-17	17	176,00	7,34	11,81	0,39	4,07	0,15	0,71	-0,42	4,34	1:2:1
Diam22-2	2	97,00	6,08	9,19	0,53	4,05	-0,27	1,19	0,39	2,16	3:1
Diam22-6	6	100,55	6,42	9,60	-1,04	5,03	-0,51	1,58	0,21	0,43	1:2:1
Diam22-17	17	177,30	6,55	11,87	0,39	2,59	0,21	0,84	-0,48	3,84	3:1
Diam28-2	2	95,40	4,21	7,62	0,29	2,55	-0,21	1,53	0,21	1,35	3:1
Diam28-17	17	176,00	6,37	10,32	0,33	3,85	0,13	0,70	-0,31	3,31	1:2:1
Diam35-2	2	95,40	4,29	7,60	0,40	1,07	-0,56	2,38	0,58	2,26	3:1
Diam35-6	6	168,00	4,15	13,22	-0,44	1,57	-0,50	2,06	0,47	1,46	3:1
Diam35-14	14	100,22	4,14	7,61	0,03	0,01	0,16	0,20	-0,72	3,87	1:1
Diam35-17	17	177,30	5,20	11,18	0,50	2,06	0,40	1,28	-0,51	2,13	3:1
Diam35-18	18	255,00	4,24	10,90	-0,35	0,98	-0,66	2,92	0,10	0,08	1:2:1

LG - Grupo de ligação

LOD - LOD Score Global

R² - porcentagem da variação fenotípica

LOD⁽²⁾ (3) (4) - valor de LOD para efeitos aditivos e de dominância

Podemos observar que as avaliações realizadas para os tempos de verão (12, 22 e 35 meses) mostraram maior número de QTLs em comparação com as avaliações para inverno. Análises para circunferência realizadas por Souza et al. (2013) mostraram maior número de QTLs em avaliações de inverno, e demonstram que outros QTLs podem ser identificados em outros grupos de ligação. Resultados do mapeamento em períodos após o verão mostraram que os diferentes QTLs identificados explicaram boa quantidade da variação fenotípica, como por exemplo, os QTLs Diam12-2 (132,00 cM), Diam22-17 (177,30 cM) e Diam35-6 (168,00 cM) com variação fenotípica (R²) de 13,68%, 11,87% e 13,22%, respectivamente.

O presente trabalho corrobora os resultados apresentados por Souza et al. (2013), que identificou pela primeira vez em seringueira, QTLs com efeitos de dominância e mostrou a importância do uso de marcadores informativos para o estudo da arquitetura genética desta característica. Foram identificados efeitos aditivos significativos para ambos os genitores em 11 QTLs, contribuindo para a variação da característica investigada, como por exemplo em Diam12-17, Diam17-2 e Diam35-18.

Os QTLs podem conter um ou vários genes relacionados com a característica estudada, sendo que a variação contínua no fenótipo representa uma ação conjunta com os efeitos ambientais (PATERSON et al., 1991). Do ponto de vista do melhoramento, os QTLs mais valiosos são aqueles identificados em múltiplos anos e que apresentam expressão positiva e consistente (CONNER; BROWN; WEEDEN, 1998). Assim, a presença de QTL no grupo de ligação 17 se torna bastante interessante, uma vez que sua detecção foi observada em todas as épocas avaliadas.

Análise cautelosa dos diferentes efeitos, posições e padrões de segregação de um QTL presente ao longo de observações continuadas são importantes, bem como alguma variação devido a presença de diferentes blocos de genes envolvidos em processos de interação genótipo x ambiente, QTLs regulados segundo a fase de desenvolvimento ou flutuações ao acaso nos dados para os diferentes períodos avaliados (CONNER; BROWN; WEEDEN, 1998). Para o grupo de ligação 2, efeitos e padrões de segregação dos QTLs nas avaliações de 17 e 35 meses, parecem indicar uma estabilidade, porém seu posicionamento precisa ser melhor entendido. Avaliações durante maior período de tempo e em diferentes locais, inclusão de maior número de marcadores nessas regiões e busca na região de interesse por genes anotados para seringueira e em espécies próximas podem ser benéficas nesse sentido.

Em termos gerais, o mapeamento de QTLs se baseia na necessidade de aumentar o conhecimento da arquitetura e herança genética de características complexas, para assim, ser possível desenvolver novas cultivares (BERNARDO, 2008). O mapeamento de QTLs para diâmetro permitiu estimar o número de QTLs e a magnitude dos efeitos aditivos e de dominância em múltiplos tempos avaliados, com condições ambientais diversas. Destacam-se nesse sentido, os QTLs identificados durante variação fenotípica no inverno, período no qual tais regiões apresentam baixa temperatura média e redução na precipitação, de forma que estes podem se tornar úteis na tentativa de se obter clones adaptados a tais condições climáticas.

Agradecimentos

Os autores agradecem ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Processos 2007/50562-4 e 2012/50491-8), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Processos

562979/2010-7, 478701/2012-8 e 162629/2015-0), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de doutorado, ao programa CAPES/Fundação Agrópolis pela bolsa de Doutorado Sanduíche desenvolvido de 10/2014 a 01/2015 e também ao Programa CAPES-Biologia Computacional.

Referências do Capítulo I

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

BERNARDO, R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. **Crop Science**, v. 48, n. 5, p. 1649–1664, 2008.

BOUTET, G. et al. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. **BMC Genomics**, v. 17, n. 1, p. 121, 2016.

BRIEN, C. **ASRemlPlus: Augments the Use of “ASReml” in Fitting Mixed Models. R package version 2.0-9**, 2016.

BUTLER, D. G. et al. **ASReml-R reference manual, release 3. Technical report** NSW Departament of Primary Industries., , 2009.

BYERS, R. L. et al. Development and mapping of SNP assays in allotetraploid cotton. **Theoretical and Applied Genetics**, v. 124, n. 7, p. 1201–1214, 2012.

CARDOSO, S. E. A. et al. Comparison of growth , yield and related traits of resistant Hevea genotypes under high South American leaf blight pressure. **Industrial Crops and Products**, v. 53, p. 337–349, 2014.

CHEN, L.; STOREY, J. D. Relaxed significance criteria for linkage analysis. **Genetics**, v. 173, n. 4, p. 2371–2381, 2006.

CHURCHILL, G. A.; DOERGE, R. W. Empirical threshold values for quantitative trait mapping. **Genetics**, v. 138, n. 3, p. 963–971, 1994.

CIRAD. **(CIRAD) Hevea Clonal Descriptions**. Disponível em: <<http://rubberclones.cirad.fr>>. Acesso em: 26 set. 2016.

CLEVENGER, J. et al. Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. **Molecular Plant**, v. 8, n. 6, p. 831–846, 2015.

CONNER, P. J.; BROWN, S. K.; WEEDEN, N. F. Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. **Theoretical and Applied Genetics**, v. 96, n. 8, p. 1027–1035, 1998.

COSTA, E. A. et al. QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. **Euphytica**, v. 211, n. 1, p. 1–16, 2016.

COVARRUBIAS-PAZARAN, G. et al. Exploiting genotyping by sequencing to characterize the genomic structure of the American cranberry through high-density linkage mapping. **BMC Genomics**, v. 17, n. 1, p. 451, 2016.

COWLING, W. A et al. Using the Animal Model to Accelerate Response to Selection in a Self-Pollinating Crop. **G3: Genes|Genomes|Genetics**, v. 5, n. 7, p. 1419–1428, 2015.

CRESTE, S.; NETO, A. T.; FIGUEIRA, A. Detection of Single Sequence Repeat Polymorphisms in Denaturing Polyacrylamide Sequencing Gels by Silver Staining. **Plant Molecular Biology Reporter**, v. 19, n. December, p. 299–306, 2001.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156–2158, 2011.

DAVEY, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature reviews. Genetics**, v. 12, n. 7, p. 499–510, jul. 2011.

DOERGE, R. W.; ZENG, Z.; WEIR, B. S. Statistical issues in the search for genes affecting quantitative traits in experimental populations. **Statistical Science**, v. 12, n. 3, p. 195–219, 1997.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v. 19, p. 11–15, 1987.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PloS one**, v. 6, n. 5, p. e19379, jan. 2011.

FENG, S. P. et al. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). **Molecular Breeding**, v. 23, n. 1, p. 85–97, 28 ago. 2009.

FIERST, J. L. Using linkage maps to correct and scaffold de novo genome assemblies : methods , challenges , and computational tools. **Frontiers in Genetics**, v. 6, n. June, p. 1–8, 2015.

GARCIA, D. et al. EST profiling of resistant and susceptible *Hevea* infected by *Microcyclus ulei*. **Physiological and Molecular Plant Pathology**, v. 76, n. 2, p. 126–136, 2011.

GAZAFFI, R. et al. A model for quantitative trait loci mapping, linkage phase, and segregation pattern estimation for a full-sib progeny. **Tree Genetics & Genomes**, p. 1–11, 23 abr. 2014.

GLAUBITZ, J. C. et al. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. **PLoS ONE**, v. 9, n. 2, 2014.

GONÇALVES, P. D. S. et al. Assessment of growth and yield performance of rubber tree clones of the IAC 500 series. **Pesquisa Agropecuária Brasileira**, v. 46, n. 12, p. 1643–1649, 2011.

GUAJARDO, V. et al. Construction of high density sweet cherry (*Prunus avium* L.) linkage maps using microsatellite markers and SNPs detected by genotyping-by-sequencing (GBS). **PLoS ONE**, v. 10, n. 5, p. 1–17, 2015.

HACKETT, C. A.; BROADFOOT, L. B. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. **Heredity**, v. 90, n. 1, p. 33–38, 2003.

HENDERSON, C. R. **Sire evaluation and genetic trends**. Proceedings of the Animal Breeding and Genetics Symposium in Honour of J.L. Lush. **Anais...**1973

JAIMES, Y. et al. Suitable climate for rubber trees affected by the South American Leaf Blight (SALB): Example for identification of escape zones in the Colombian middle Magdalena. **Crop Protection**, v. 81, p. 99–114, 2016.

KOSAMBI, D. D. the Estimation of Map Distances From Recombination Values. **Annals of Eugenics**, v. 12, p. 172–175, 1943.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods**, v. 9, n. 4, p. 357–359, 2012.

LAU, N. et al. The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. **Nature Publishing Group**, n. January, p. 1–14, 2016.

LE GUEN, V. et al. A rubber tree's durable resistance to *Microcyclus ulei* is conferred by a qualitative gene and a major quantitative resistance factor. **Tree Genetics and Genomes**, v. 7, n. 5, p. 877–889, 2011a.

LE GUEN, V. et al. Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*). **Plant Breeding**, v. 130, n. 2, p. 294–296, 1 abr. 2011b.

LESPINASSE, D. et al. A saturated genetic linkage map of rubber tree (*Hevea* spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. **Theoretical and Applied Genetics**, v. 100, p. 127–138, 2000.

LU, F. et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. **PLoS genetics**, v. 9, n. 1, p. e1003215, jan. 2013.

MANTELO, C. C. et al. Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species. **BMC research notes**, v. 5, p. 329, jan. 2012.

MANTELLO, C. C. et al. De Novo assembly and transcriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. **PLoS ONE**, v. 9, n. 7, p. 1–14, 2014.

MARGARIDO, G. R. A.; SOUZA, A. P.; GARCIA, A. A. F. OneMap: software for genetic mapping in outcrossing species. **Hereditas**, v. 144, p. 78–79, 2007.

MARUBODEE, R. et al. Construction of an SSR and RAD-marker based molecular linkage map of *Vigna vexillata* (L.) A. Rich. **PLoS ONE**, v. 10, n. 9, p. 1–16, 2015.

MCCALLUM, S. et al. Construction of a SNP and SSR linkage map in autotetraploid blueberry using genotyping by sequencing. **Molecular Breeding**, v. 36, n. 4, p. 1–24, 2016.

MRODE, R. A. **Mrode, R.A. Linear models for the prediction of animal breeding values**. Cabi: [s.n.].

NIRAPATHPONGPORN, K. et al. Development and mapping of functional expressed sequence tag-derived simple sequence repeat markers in a rubber tree RRIM600 × PB217 population. **Molecular Breeding**, v. 36, n. 4, p. 39, 2016.

NOVALINA, A. D. S. Construction of *Hevea brasiliensis* Genetic Linkage Map and Identification of Quantitative Trait Loci Using RAPD Markers. **International Journal on Advanced Science, Engineering and Information Technology**, v. 3, n. 1, p. 71, 2013.

OBERACHER, H. On the use of different mass spectrometric techniques for characterization of sequence variability in genomic DNA. **Analytical and Bioanalytical Chemistry**, v. 391, n. 1, p. 135–149, 2008.

PAGET, M. F. et al. Trial heterogeneity and variance models in the genetic evaluation of potato tuber yield. **Plant Breeding**, v. 134, n. 2, p. 203–211, 2015.

PATERSON, A. H. et al. Mendelian factors underlying quantitative traits in tomato: Comparison across species, generations, and environments. **Genetics**, v. 127, n. 1, p. 181–197, 1991.

PINHEIRO, E.; PINHEIRO, F. S. V. Heveicultura em área de escape. In: ALVARENGA, A.P., CARMO, C. A. F. S. (Ed.). **Seringueira**. 1ª. ed. Viçosa: Epamig, 2008. p. 83–126.

POOTAKHAM, W. et al. Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). **Frontiers in plant science**, v. 6, n. May, p. 367, 2015.

PRIYADARSHAN, P. M.; CLÉMENT-DEMANGE, A. Breeding *Hevea* Rubber: Formal and Molecular Genetics. **Advances in Genetics**, v. 52, n. February 2004, p. 51–115, 2004.

R CORE TEAM. R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>, v. 55, p. 275–286, 2015.

RAHMAN, A. Y. A. et al. Draft genome sequence of the rubber tree *Hevea brasiliensis*. **BMC genomics**, v. 14, n. January, p. 75, jan. 2013.

RAO, G. P.; KOLE, P. C. Evaluation of Brazilian wild *Hevea* germplasm for cold tolerance: genetic variability in the early mature growth. **Journal of Forestry Research**, v. 27, n. 4, p. 755–765, 2016.

SAHA, T.; PRIYADARSHAN, P. M. Genomics of *Hevea* Rubber. In: SCHNELL, R. J.; PRIYADARSHAN, P. M. (Eds.). . **Genomics of Tree Crops**. New York, NY: Springer New York, 2012. p. 261–298.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978.

SHEARMAN, J. R. et al. Assembly and analysis of a male sterile rubber tree mitochondrial genome reveals DNA rearrangement events and a novel transcript. **BMC plant biology**, v. 14, n. 1, p. 45, jan. 2014.

SHEARMAN, J. R. et al. SNP identification from RNA sequencing and linkage map construction of rubber tree for anchoring the draft genome. **PLoS ONE**, v. 10, n. 4, p. 1–12, 2015.

SHIELDS, D. C. et al. Error filtration, interference, and the human linkage map. **Proceedings of the National Academy of Sciences of the United States of America**, v. 88, n. 15, p. 6501–6505, 1991.

SILPI, U. et al. Effect of tapping activity on the dynamics of radial growth of *Hevea brasiliensis* trees. **Tree physiology**, v. 26, n. 12, p. 1579–1587, 2006.

SILVA, C. C. et al. Leaf-, panel- and latex-expressed sequenced tags from the rubber tree (*Hevea brasiliensis*) under cold-stressed and suboptimal growing conditions: the development of gene-targeted functional markers for stress response. **Molecular breeding : new strategies in plant improvement**, v. 34, n. 3, p. 1035–1053, 29 jan. 2014a.

SILVA, L. G. et al. Performance of different rubber tree clones against South American leaf blight (*Microcyclus ulei*). **Forest Pathology**, v. 44, p. 211–218, 2014b.

SMITH, A.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: na overview of current mixed model approaches. **The Journal of Agricultural Science**, v. 143, p. 449–462, 2005.

SOUZA, L. M. et al. Microsatellites from rubber tree (*Hevea brasiliensis*) for genetic diversity analysis and cross-amplification in six *Hevea* wild species. **Conservation Genetics Resources**, v. 1, n. 1, p. 75–79, 2009.

SOUZA, L. M. et al. QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (*Hevea brasiliensis*) Evaluated in a Sub-Tropical Climate. **PLoS ONE**, v. 8, n. 4, p. e61238, jan. 2013.

SOUZA, L. M. et al. Development of single nucleotide polymorphism markers in the large and complex rubber tree genome using next-generation sequence data. **Molecular Breeding**, v. 36, n. 115, p. 1–10, 2016.

TANG, C. et al. The rubber tree genome reveals new insights into rubber production and species adaptation. **Nature Plants**, v. 2, n. 6, p. 16073, 2016.

TRIWITAYAKORN, K. et al. Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. **DNA research : an international journal for rapid publication of reports on genes and genomes**, v. 18, n. 6, p. 471–82, dez. 2011.

VOORRIPS, R. E. MapChart: software for the graphical presentation of linkage maps and QTLs. **The Journal of heredity**, v. 93, n. 1, p. 77–78, 2002.

WANG, J. et al. High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. **BMC genomics**, v. 10, p. 561, 2009.

WARD, J. A et al. Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. **BMC genomics**, v. 14, p. 2, 2013.

WU, R. et al. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. **Theoretical population biology**, v. 61, n. 3, p. 349–63, maio 2002.

XU, S. Quantitative trait locus mapping can benefit from segregation distortion. **Genetics**, v. 180, n. 4, p. 2201–2208, 2008.

YANG, S. et al. The first genetic map of a synthesized allohexaploid Brassica with A , B and C genomes based on simple sequence repeat markers. **Theoretical and applied genetics**, v. 129, p. 689–701, 2016.

ZENG, Z. B. Precision mapping of quantitative trait loci. **Genetics**, v. 136, n. 4, p. 1457–1468, 1994.

Conclusões

- O polimorfismo de marcadores, em especial dos microssatélites, possibilitou a construção de um mapa genético saturado e representando os 18 grupos de ligação de seringueira.
- A genotipagem da população de mapeamento por meio da abordagem de GBS permitiu a geração de grande quantidade de dados, e possibilitou a saturação de regiões cromossômicas que não haviam sido saturadas utilizando outros tipos de marcadores moleculares.
- Os dados gerados pela técnica de GBS são dinâmicos e poderão ser reanalisados para melhor aproveitamento dos dados sempre que uma nova abordagem de análise estatística ou sequenciamento genômico surgirem.
- QTLs com efeitos aditivos e de dominância foram detectados, com predominância de efeitos aditivos detectados no genitor GT1.
- O mapa genético saturado e o mapeamento de QTL forneceram simultaneamente novas informações para a espécie e ferramentas para novas abordagens de investigação, bem como disponibilizaram importantes informações aos programas de melhoramento genético.

Perspectivas

O papel marcante dos marcadores microsatélites relacionado ao seu polimorfismo também avaliados no presente trabalho, evidencia que a importância dessa classe de marcadores continuará também na era genômica. A quantidade de marcadores SSRs informativos para a população bastante inferior em relação aos SNPs de GBS, tornou o processo de genotipagem manual factível, chegando a mostrar vantagens quanto ao menor número de genótipos incorretos produzidos. Outros trabalhos envolvendo essa classe de marcadores podem surgir para aproveitar tais características. Naturalmente, o processo de avaliação da aplicabilidade dessa classe de marcadores não é fácil, necessitando avaliar a quantidade de marcadores a serem utilizados.

O sucesso na construção do mapa genético para *Hevea brasiliensis* é resultado da utilização de marcadores SNPs da metodologia de GBS e sua utilização a partir de um mapa com boa quantidade de marcadores pré-ordenados. É importante notar ainda, a grande contribuição da presença de um genoma de referência, bem como a utilização de filtros adequados para obtenção final dos SNPs úteis ao mapeamento. Dessa forma, os marcadores moleculares mapeados no presente trabalho e também aqueles que não foram aproveitados e/ou avaliados, sendo ainda disponíveis na literatura, podem ser utilizados novamente em conjunto com novas metodologias que venham a surgir, para que assim, se possa ter outras informações sobre o complexo genoma de seringueira.

A possibilidade de aumentar a quantidade de dados fenotípicos poderá contribuir ainda mais para a detecção e caracterização de regiões relacionadas a características quantitativas. A população de mapeamento utilizada tem especial importância, pois se trata da primeira população de seringueira construída a partir de uma parceria de órgãos públicos (Universidade Estadual de Campinas e Instituto Agrônomo de Campinas), de forma que a coleta de dados está apenas em seu início, e pode resultar em novos avanços tecnológicos para a produção de borracha.

Referências Gerais

- AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.
- BAIRD, N. A et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. **PloS one**, v. 3, n. 10, p. e3376, jan. 2008.
- BENCHIMOL, L. L.; SOUZA JR, C. L.; SOUZA, A. P. Microsatellite-assisted backcross selection in maize. **Genetics and Molecular Biology**, v. 28, n. 4, p. 789–797, 2005.
- BERNARDO, R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. **Crop Science**, v. 48, n. 5, p. 1649–1664, 2008.
- BOUTET, G. et al. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. **BMC Genomics**, v. 17, n. 1, p. 121, 2016.
- BRADSHAW, J. E. et al. QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*). **Theoretical and Applied Genetics**, v. 116, n. 2, p. 193–211, 2008.
- BRIEN, C. **ASRemlPlus: Augments the Use of “ASReml” in Fitting Mixed Models. R package version 2.0-9**, 2016.
- BUTLER, D. G. et al. **ASReml-R reference manual, release 3. Technical report** NSW Department of Primary Industries., , 2009.
- BYERS, R. L. et al. Development and mapping of SNP assays in allotetraploid cotton. **Theoretical and Applied Genetics**, v. 124, n. 7, p. 1201–1214, 2012.
- CAMARGO, A. P.; CAMARGO, M. B. P. Aptidão Climática da heveicultura no Brasil. In: ALVARENGA, A.P., CARMO, C. A. F. S. (Ed.). **Seringueira**. 1^a. ed. Viçosa: Epamig, 2008. p. 25–50.
- CARDOSO, S. E. A. et al. Comparison of growth , yield and related traits of resistant Hevea genotypes under high South American leaf blight pressure. **Industrial Crops and Products**, v. 53, p. 337–349, 2014.
- CARNEIRO, M. S.; VIEIRA, M. L. C. Mapas genéticos em plantas. **Bragantia**, v. 61, n. 2, p. 89–100, 2002.
- CENTURION, M. A. P. C. et al. Efeito do manejo da entrelinha da seringueira sobre as propriedades químicas do solo , o estado nutricional e o crescimento. **Revista Árvore**, v. 29, n. 2, p. 185–193, 2005.
- CHAN, A. P. et al. Draft genome sequence of the oilseed species *Ricinus communis*. **Nature biotechnology**, v. 28, n. 9, p. 951–6, 2010.

CHARCOSSET, A.; MOREAU, L. Use of molecular markers for the development of new cultivars and the evaluation of genetic diversity. **Euphytica**, v. 137, n. 1, p. 81–94, 2004.

CHEN, L.; STOREY, J. D. Relaxed significance criteria for linkage analysis. **Genetics**, v. 173, n. 4, p. 2371–2381, 2006.

CHOW, K. S. et al. Metabolic routes affecting rubber biosynthesis in *Hevea brasiliensis* latex. **Journal of Experimental Botany**, v. 63, n. 5, p. 1863–1871, 2012.

CHOW, K.-S. et al. Insights into rubber biosynthesis from transcriptome analysis of *Hevea brasiliensis* latex. **Journal of experimental botany**, v. 58, n. 10, p. 2429–40, jan. 2007.

CHURCHILL, G. A.; DOERGE, R. W. Empirical threshold values for quantitative trait mapping. **Genetics**, v. 138, n. 3, p. 963–971, 1994.

CIRAD. **(CIRAD) Hevea Clonal Descriptions**. Disponível em: <<http://rubberclones.cirad.fr>>. Acesso em: 26 set. 2016.

CLEVENGER, J. et al. Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. **Molecular Plant**, v. 8, n. 6, p. 831–846, 2015.

COLLARD, B. C. Y.; MACKILL, D. J. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. **Philosophical Transactions of the Royal Society**, v. 363, n. August 2007, p. 557–572, 2008.

CONNER, P. J.; BROWN, S. K.; WEEDEN, N. F. Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. **Theoretical and Applied Genetics**, v. 96, n. 8, p. 1027–1035, 1998.

CONSON, A. R. O. et al. Genetic structure of the Atlantic Rainforest tree species *Luehea divaricata* (Malvaceae). **Genetica**, v. 141, n. 4-6, p. 205–15, jun. 2013.

COSTA, E. A. et al. QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. **Euphytica**, v. 211, n. 1, p. 1–16, 2016.

COVARRUBIAS-PAZARAN, G. et al. Exploiting genotyping by sequencing to characterize the genomic structure of the American cranberry through high-density linkage mapping. **BMC Genomics**, v. 17, n. 1, p. 451, 2016.

COWLING, W. A et al. Using the Animal Model to Accelerate Response to Selection in a Self-Pollinating Crop. **G3: Genes|Genomes|Genetics**, v. 5, n. 7, p. 1419–1428, 2015.

CRESTE, S.; NETO, A. T.; FIGUEIRA, A. Detection of Single Sequence Repeat Polymorphisms in Denaturing Polyacrylamide Sequencing Gels by Silver Staining. **Plant Molecular Biology Reporter**, v. 19, n. December, p. 299–306, 2001.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156–2158, 2011.

DAVEY, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature reviews. Genetics**, v. 12, n. 7, p. 499–510, jul. 2011.

DE DONATO, M. et al. Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. **PLoS ONE**, v. 8, n. 5, p. 1–9, 2013.

DOERGE, R. W. Mapping and analysis of quantitative trait loci in experimental populations. **Nature reviews. Genetics**, v. 3, n. 1, p. 43–52, jan. 2002.

DOERGE, R. W.; ZENG, Z.; WEIR, B. S. Statistical issues in the search for genes affecting quantitative traits in experimental populations. **Statistical Science**, v. 12, n. 3, p. 195–219, 1997.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v. 19, p. 11–15, 1987.

EDWARDS, D.; BATLEY, J.; SNOWDON, R. J. Accessing complex crop genomes with next-generation sequencing. **Theoretical and Applied Genetics**, v. 126, p. 1–11, 2013.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PloS one**, v. 6, n. 5, p. e19379, jan. 2011.

FENG, S. P. et al. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). **Molecular Breeding**, v. 23, n. 1, p. 85–97, 28 ago. 2009.

FERREIRA, M. E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. Brasília: Embrapa - Cenargem, 1996.

FIERST, J. L. Using linkage maps to correct and scaffold de novo genome assemblies : methods , challenges , and computational tools. **Frontiers in Genetics**, v. 6, n. June, p. 1–8, 2015.

GARCIA, D. et al. EST profiling of resistant and susceptible *Hevea* infected by *Microcyclus ulei*. **Physiological and Molecular Plant Pathology**, v. 76, n. 2, p. 126–136, 2011.

GAZAFFI, R. et al. A model for quantitative trait loci mapping, linkage phase, and segregation pattern estimation for a full-sib progeny. **Tree Genetics & Genomes**, p. 1–11, 23 abr. 2014.

GLAUBITZ, J. C. et al. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. **PLoS ONE**, v. 9, n. 2, 2014.

GONÇALVES, P. D. S. et al. Assessment of growth and yield performance of rubber tree clones of the IAC 500 series. **Pesquisa Agropecuária Brasileira**, v. 46, n. 12, p. 1643–1649, 2011.

GONÇALVES, P. DE S.; CARDOSO, M.; ORTOLANI, A. A. Origem, variabilidade e domesticação de Hevea; uma revisão. **Pesquisa Agropecuária Brasileira**, v. 25, p. 135–156, 1990.

GONÇALVES, P. S.; FONTES, J. R. A. Domesticação e Melhoramento da Seringueira. In: Borém A, Lopes MTG, Clement CR & Noda H (Eds). Domesticação e Melhoramento: Espécies Amazônicas. **Viçosa**, p. 395–423, 2009.

GONÇALVES, P. S.; MARQUES, J. R. B. Melhoramento Genético da Seringueira: Passado, Presente e Futuro. In: ALVARENGA, A. P.; CARMO, C. A. F. S. (Eds.). **Seringueira**. 1ª. ed. Viçosa: Epamig, 2008. p. 399–498.

GRATTAPAGLIA, D.; SEDEROFF, R. Genetic Linkage Maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a Pseudo-Testcross: Mapping Strategy and RAPD markers. **Genetics**, v. 1137, p. 1121–1137, 1994.

GUAJARDO, V. et al. Construction of high density sweet cherry (*Prunus avium* L.) linkage maps using microsatellite markers and SNPs detected by genotyping-by-sequencing (GBS). **PLoS ONE**, v. 10, n. 5, p. 1–17, 2015.

HACKETT, C. A; BROADFOOT, L. B. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. **Heredity**, v. 90, n. 1, p. 33–38, 2003.

HAYWARD, A. et al. SNP discovery and applications in *Brassica napus*. **Journal of Plant Biotechnology**, v. 39, n. 1, p. 49–61, 31 mar. 2012.

HENDERSON, C. R. **Sire evaluation and genetic trends**. Proceedings of the Animal Breeding and Genetics Symposium in Honour of J.L. Lush. **Anais...**1973

HODEL, R. G. J. et al. The report of my death was an exaggeration: a review for researches using microsatellite in the 21th century. **Applications in Plant Sciences**, v. 4, n. 6, p. 1–13, 2016.

IPEF. (IPEF) **Instituto de Pesquisa e Estudos Florestais. Hevea brasiliensis (seringueira)**. Disponível em: <<http://www.ipef.br/identificacao/hevea.brasiliensi>>. Acesso em: 10 ago. 2016.

JAIMES, Y. et al. Suitable climate for rubber trees affected by the South American Leaf Blight (SALB): Example for identification of escape zones in the Colombian middle Magdalena. **Crop Protection**, v. 81, p. 99–114, 2016.

JANSEN, R. C.; STAM, P. High resolution of quantitative traits into multiple loci via interval mapping. **Genetics**, v. 136, n. 4, p. 1447–1455, 1994.

- KAO, C. H.; ZENG, Z. B.; TEASDALE, R. D. Multiple interval mapping for quantitative trait loci. **Genetics**, v. 152, n. 3, p. 1203–16, 1999.
- KING, A. J. et al. Identification of QTL markers contributing to plant growth, oil yield and fatty acid composition in the oilseed crop *Jatropha curcas* L. **Biotechnology for Biofuels**, v. 8, n. 1, p. 160, 2015.
- KOSAMBI, D. D. the Estimation of Map Distances From Recombination Values. **Annals of Eugenics**, v. 12, p. 172–175, 1943.
- LANDER, E. S.; BOTSTEIN, D. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. **Genetics**, v. 121, p. 185–199, 1989.
- LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods**, v. 9, n. 4, p. 357–359, 2012.
- LAU, N. et al. The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. **Nature Publishing Group**, n. January, p. 1–14, 2016.
- LE GUEN, V. et al. Long lasting rubber tree resistance to *Microcyclus ulei* characterized by reduced conidial emission and absence of teleomorph. **Crop Protection**, v. 27, p. 1498–1503, 2008.
- LE GUEN, V. et al. Genetic structure of Amazonian populations of *Hevea brasiliensis* is shaped by hydrographical network and isolation by distance. **Tree Genetics & Genomes**, v. 5, n. 4, p. 673–683, 23 maio 2009.
- LE GUEN, V. et al. A rubber tree's durable resistance to *Microcyclus ulei* is conferred by a qualitative gene and a major quantitative resistance factor. **Tree Genetics and Genomes**, v. 7, n. 5, p. 877–889, 2011a.
- LE GUEN, V. et al. Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*). **Plant Breeding**, v. 130, n. 2, p. 294–296, 1 abr. 2011b.
- LEE, H. C. et al. Bioinformatics tools and databases for analysis of next-generation sequence data. **Briefings in functional genomics**, v. 11, n. 1, p. 12–24, jan. 2011.
- LEITCH, A. R. et al. Molecular cytogenetic studies in rubber , *Hevea brasiliensis* Muell. Arg. (Euphorbiaceae). **Genome**, v. 41, p. 464–467, 1998.
- LESPINASSE, D. et al. A saturated genetic linkage map of rubber tree (*Hevea* spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. **Theoretical and Applied Genetics**, v. 100, p. 127–138, 2000.
- LI, Y.; HE, M. Genetic Mapping and QTL Analysis of Growth-Related Traits in *Pinctada fucata* Using Restriction-Site Associated DNA Sequencing. **PLoS ONE**, v. 9, n. 11, 2014.

LIEBEREI, R. South American Leaf Blight of the Rubber Tree (*Hevea* spp.): New Steps in Plant Domestication using Physiological Features and Molecular Markers. **Annals of botany**, v. 100, p. 1125–1142, 2007.

LU, F. et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. **PLoS genetics**, v. 9, n. 1, p. e1003215, jan. 2013.

MANTELO, C. C. et al. Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species. **BMC research notes**, v. 5, p. 329, jan. 2012.

MANTELO, C. C. et al. De Novo assembly and transcriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. **PLoS ONE**, v. 9, n. 7, p. 1–14, 2014.

MARGARIDO, G. R. A.; SOUZA, A. P.; GARCIA, A. A. F. OneMap: software for genetic mapping in outcrossing species. **Hereditas**, v. 144, p. 78–79, 2007.

MARUBODEE, R. et al. Construction of an SSR and RAD-marker based molecular linkage map of *Vigna vexillata* (L.) A. Rich. **PLoS ONE**, v. 10, n. 9, p. 1–16, 2015.

MCCALLUM, S. et al. Construction of a SNP and SSR linkage map in autotetraploid blueberry using genotyping by sequencing. **Molecular Breeding**, v. 36, n. 4, p. 1–24, 2016.

MCCOUCH, S. R. et al. Microsatellite marker development, mapping and applications in rice genetics and breeding. **Plant Molecular Biology**, v. 35, p. 89–99, 1997.

MOOIBROEK, H.; CORNISH, K. Alternative sources of natural rubber. **Applied microbiology and biotechnology**, v. 53, n. 4, p. 355–65, abr. 2000.

MORI, G. M.; ZUCCHI, M. I.; SOUZA, A. P. Multiple-Geographic-Scale Genetic Structure of Two Mangrove Tree Species: The Roles of Mating System, Hybridization, Limited Dispersal and Extrinsic Factors. **PloS one**, p. 1–23, 2015.

MRODE, R. A. **Mrode, R.A. Linear models for the prediction of animal breeding values**. Cabi: [s.n.].

NIRAPATHPONGPORN, K. et al. Development and mapping of functional expressed sequence tag-derived simple sequence repeat markers in a rubber tree RRIM600 × PB217 population. **Molecular Breeding**, v. 36, n. 4, p. 39, 2016.

NOVALINA, A. D. S. Construction of *Hevea brasiliensis* Genetic Linkage Map and Identification of Quantitative Trait Loci Using RAPD Markers. **International Journal on Advanced Science, Engineering and Information Technology**, v. 3, n. 1, p. 71, 2013.

OBERACHER, H. On the use of different mass spectrometric techniques for characterization of sequence variability in genomic DNA. **Analytical and Bioanalytical Chemistry**, v. 391, n. 1, p. 135–149, 2008.

OLIVEIRA, K. M. et al. Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp .) commercial cross. **Molecular Breeding**, v. 20, p. 189–208, 2007.

ONG, S. N. **Chromosome morphology at pachytene stage in *Hevea brasiliensis*: a preliminar report**. Sripathi B(ed) Proceedings of the international rubber conference. **Anais...**Kuala Lumpur: Rubber Research Institute of Malaysia, 1975

PAGET, M. F. et al. Trial heterogeneity and variance models in the genetic evaluation of potato tuber yield. **Plant Breeding**, v. 134, n. 2, p. 203–211, 2015.

PAIVA, J. R. **Variabilidade enzimática em populações naturais de seringueira**. [s.l.] ESALQ-USP, 1992.

PATERSON, A. H. et al. Mendelian factors underlying quantitative traits in tomato: Comparison across species, generations, and environments. **Genetics**, v. 127, n. 1, p. 181–197, 1991.

PEEL, M. C.; FINLAYSON, B. L.; MCMAHON, T. A. Updated world map of the Köppen-Geiger climate classification. **Hydrology and Earth System Science**, v. 11, n. 5, p. 1633–1644, 2007.

PINHEIRO, E.; PINHEIRO, F. S. V. Heveicultura em área de escape. In: ALVARENGA, A.P., CARMO, C. A. F. S. (Ed.). **Seringueira**. 1ª. ed. Viçosa: Epamig, 2008. p. 83–126.

PIRES, J. M.; SECCO, R. S.; GOMES, J. I. **Taxonomia e fitogeografia das seringueiras (*Hevea* spp)**. Belém: Embrapa Amazônia Oriental, 2002.

POLAND, J. A; RIFE, T. W. Genotyping-by-Sequencing for Plant Breeding and Genetics. **Plant Genome**, v. 5, n. November, p. 92–102, 2012.

POOTAKHAM, W. et al. Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). **Frontiers in plant science**, v. 6, n. May, p. 367, 2015.

PRIYADARSHAN, P. M.; CLÉMENT-DEMANGE, A. Breeding Hevea Rubber: Formal and Molecular Genetics. **Advances in Genetics**, v. 52, n. February 2004, p. 51–115, 2004.

PRIYADARSHAN, P. M.; GONCALVES, P. DE S. Hevea gene pool for breeding. **Genetic Resources and Crop Evolution**, v. 50, p. 101–114, 2003.

R CORE TEAM. R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>, v. 55, p. 275–286, 2015.

RAHMAN, A. Y. A. et al. Draft genome sequence of the rubber tree *Hevea brasiliensis*. **BMC genomics**, v. 14, n. January, p. 75, jan. 2013.

RAMAER, H. Cytology in *Hevea*. **Genetica**, v. 17, p. 193–236, 1935.

RAO, G. P.; KOLE, P. C. Evaluation of Brazilian wild *Hevea* germplasm for cold tolerance: genetic variability in the early mature growth. **Journal of Forestry Research**, v. 27, n. 4, p. 755–765, 2016.

RIVANO, F. et al. Assessing resistance of rubber tree clones to *Microcyclus ulei* in large-scale clone trials in Ecuador: A less time-consuming field method. **European Journal of Plant Pathology**, v. 126, n. 4, p. 541–552, 2010.

SAHA, T.; PRIYADARSHAN, P. M. Genomics of *Hevea* Rubber. In: SCHNELL, R. J.; PRIYADARSHAN, P. M. (Eds.). **Genomics of Tree Crops**. New York, NY: Springer New York, 2012. p. 261–298.

SANTOS, E. S. L.; CERQUEIRA-SILVA, C. B. M.; MORI, G. M. Genetic Structure and Molecular Diversity of Cacao Plants Established as Local Varieties for More than Two Centuries: The Genetic History of Cacao Plantations in Bahia, Brazil. **PloS one**, p. 1–18, 2015.

SCHLÖTTERER, C. The evolution of molecular markers — just a matter of fashion? **Nature Reviews Genetics**, v. 5, p. 63–69, 2004.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978.

SECCO, R. S. A Botânica da seringueira. In: ALVARENGA, A.P., CARMO, C. A. F. S. (Ed.). **Seringueira**. 1ª. ed. Viçosa: Epamig, 2008. p. 1–24.

SHEARMAN, J. R. et al. Assembly and analysis of a male sterile rubber tree mitochondrial genome reveals DNA rearrangement events and a novel transcript. **BMC plant biology**, v. 14, n. 1, p. 45, jan. 2014.

SHEARMAN, J. R. et al. SNP identification from RNA sequencing and linkage map construction of rubber tree for anchoring the draft genome. **PLoS ONE**, v. 10, n. 4, p. 1–12, 2015.

SHIELDS, D. C. et al. Error filtration, interference, and the human linkage map. **Proceedings of the National Academy of Sciences of the United States of America**, v. 88, n. 15, p. 6501–6505, 1991.

SILPI, U. et al. Effect of tapping activity on the dynamics of radial growth of *Hevea brasiliensis* trees. **Tree physiology**, v. 26, n. 12, p. 1579–1587, 2006.

SILVA, C. C. et al. Leaf-, panel- and latex-expressed sequenced tags from the rubber tree (*Hevea brasiliensis*) under cold-stressed and suboptimal growing conditions: the development of gene-targeted functional markers for stress response. **Molecular**

breeding : new strategies in plant improvement, v. 34, n. 3, p. 1035–1053, 29 jan. 2014a.

SILVA, L. G. et al. Performance of different rubber tree clones against South American leaf blight (*Microcyclus ulei*). **Forest Pathology**, v. 44, p. 211–218, 2014b.

SMITH, A.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. **The Journal of Agricultural Science**, v. 143, p. 449–462, 2005.

SOUZA, L. M. et al. Microsatellites from rubber tree (*Hevea brasiliensis*) for genetic diversity analysis and cross-amplification in six *Hevea* wild species. **Conservation Genetics Resources**, v. 1, n. 1, p. 75–79, 2009.

SOUZA, L. M. et al. QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (*Hevea brasiliensis*) Evaluated in a Sub-Tropical Climate. **PLoS ONE**, v. 8, n. 4, p. e61238, jan. 2013.

SOUZA, L. M. et al. Genetic Diversity Strategy for the Management and Use of Rubber Genetic Resources: More than 1 , 000 Wild and Cultivated Accessions in a 100-Genotype Core Collection. **PloS one**, p. 1–20, 2015.

SOUZA, L. M. et al. Development of single nucleotide polymorphism markers in the large and complex rubber tree genome using next-generation sequence data. **Molecular Breeding**, v. 36, n. 115, p. 1–10, 2016.

TANG, C. et al. The rubber tree genome reveals new insights into rubber production and species adaptation. **Nature Plants**, v. 2, n. 6, p. 16073, 2016.

TANKSLEY, S. D. Comparative linkage maps of the rice and maize genomes. **Genomics**, v. 90, p. 7980–7984, 1993.

TRIWITAYAKORN, K. et al. Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. **DNA research : an international journal for rapid publication of reports on genes and genomes**, v. 18, n. 6, p. 471–82, dez. 2011.

VAN ORSOUW, N. J. et al. Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. **PLoS ONE**, v. 2, n. 11, 2007.

VOORRIPS, R. E. MapChart: software for the graphical presentation of linkage maps and QTLs. **The Journal of heredity**, v. 93, n. 1, p. 77–78, 2002.

VOS, P. et al. AFLP: A new technique for DNA fingerprinting. **Nucleic Acids Research**, v. 23, n. 21, p. 4407–4414, 1995.

WANG, J. et al. High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. **BMC genomics**, v. 10, p. 561, 2009.

WANG, W. et al. Cassava genome from a wild ancestor to cultivated varieties. **Nature communications**, v. 5, p. 5110, 2014.

WARD, J. A et al. Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. **BMC genomics**, v. 14, p. 2, 2013.

WU, R. et al. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. **Theoretical population biology**, v. 61, n. 3, p. 349–63, maio 2002.

WYCHERLEY, P. R. Rubber (*Hevea brasiliensis* Euphorbeaceae). In: N.W.SIMMONDS (Ed.). . **Evolution of crop plants**. London: Longman Group Ltd, 1976. p. 77–80.

XIA, Z. et al. RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. **Plant molecular biology**, v. 77, n. 3, p. 299–308, out. 2011.

XU, S. Quantitative trait locus mapping can benefit from segregation distortion. **Genetics**, v. 180, n. 4, p. 2201–2208, 2008.

YANG, H. et al. Sequencing consolidates molecular markers with plant breeding practice. **Theoretical and Applied Genetics**, v. 128, p. 779–795, 2015.

YANG, S. et al. The first genetic map of a synthesized allohexaploid Brassica with A , B and C genomes based on simple sequence repeat markers. **Theoretical and applied genetics**, v. 129, p. 689–701, 2016.

ZANE, L.; BARGELLONI, L.; PATARNELLO, T. Strategies for microsatellite isolation : a review. **Molecular Ecology**, v. 11, p. 1–16, 2002.

ZENG, Z. B. Precision mapping of quantitative trait loci. **Genetics**, v. 136, n. 4, p. 1457–1468, 1994.

ZUREK, P. R.; TOPP, C. N.; BENFEY, P. N. Quantitative Trait Locus Mapping Reveals Regions of the Maize Genome Controlling Root System Architecture. **Plant Physiology**, v. 167, n. 4, p. 1487–1496, 2015.

ANEXOS

Anexo 1



RESEARCH ARTICLE

Genetic Diversity Strategy for the Management and Use of Rubber Genetic Resources: More than 1,000 Wild and Cultivated Accessions in a 100-Genotype Core Collection



CrossMark
click for updates

Livia Moura de Souza^{1*}, Vincent Le Guen^{2*}, Carlos Bernardo Moreno Cerqueira-Silva^{1,3}, Carla Cristina Silva¹, Camila Campos Mantello¹, Andre Ricardo Oliveira Conson¹, João Paulo Gomes Vianna^{1,4}, Maria Imaculada Zucchi⁴, Erivaldo José Scaloppi Junior⁷, Josefino de Freitas Fialho⁵, Mario Luis Teixeira de Moraes⁶, Paulo de Souza Gonçalves⁷, Anete Pereira de Souza^{1,8*}

1 Molecular Biology and Genetic Engineering Center (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil, **2** Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) UMR AGAP, Montpellier, Hérault, France, **3** Laboratory of Applied Molecular Genetics, Department of Exact and Natural Sciences, State University of Southwest Bahia (UESB), Itapetinga, BA, Brazil, **4** Agência Paulista de Tecnologia dos Agronegócios (APTA), Pólo Regional Noroeste Paulista, Votuporanga, SP, Brazil, **5** Embrapa Cerrados (EMBRAPA), Planaltina, DF, Brazil, **6** Departamento de Fitotecnia, Faculdade de Engenharia de Ilha Solteira (UNESP)—Universidade Estadual Paulista, Ilha Solteira, SP, Brazil, **7** Rubber Research Advanced Center (CAPSA), Agronomical Institute (IAC), Votuporanga, SP, Brazil, **8** Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP) UNICAMP, Campinas, SP, Brazil

* These authors contributed equally to this work.

* anete@unicamp.br

OPEN ACCESS

Citation: de Souza LM, Le Guen V, Cerqueira-Silva CBM, Silva CC, Mantello CC, Conson ARQ, et al. (2015) Genetic Diversity Strategy for the Management and Use of Rubber Genetic Resources: More than 1,000 Wild and Cultivated Accessions in a 100-Genotype Core Collection. PLoS ONE 10(7): e0134607. doi:10.1371/journal.pone.0134607

Editor: Roberto Papa, Università Politecnica delle Marche, ITALY

Received: February 19, 2015

Accepted: July 12, 2015

Published: July 30, 2015

Copyright: © 2015 de Souza et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors gratefully acknowledge the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) (2007/50392-1; 2012/50491-8) for financial support and scholarships to CCM (2011/50188-0), CCS (2009/52975-0), and LMS (2012/05473-1), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support (478701/2012-8; 402954/2012-2) and research fellowships to APS and PSG, and the

Abstract

The rubber tree [*Hevea brasiliensis* (Willd. ex A. de Juss.) Muell. Arg.] is the only plant species worldwide that is cultivated for the commercial production of natural rubber. This study describes the genetic diversity of the *Hevea* spp. complex that is available in the main *ex situ* collections of South America, including Amazonian populations that have never been previously described. Genetic data were analyzed to determine the genetic structure of the wild populations, quantify the allelic diversity and suggest the composition of a core collection to capture the maximum genetic diversity within a minimal sample size. A total of 1,117 accessions were genotyped with 13 microsatellite markers. We identified a total of 408 alleles, 319 of which were shared between groups and 89 that were private in different groups of accessions. In a population structure and principal component analysis, the level of clustering reflected a primary division into the following two subgroups: cluster 1, which consisted of varieties from the advanced breeding germplasm that originated from the Wickham and Mato Grosso accessions; and cluster 2, which consisted of the wild germplasm from the Acre, Amazonas, Pará and Rondônia populations and *Hevea* spp. The analyses revealed a high frequency of gene flow between the groups, with the genetic differentiation coefficient (G_{ST}) estimated to be 0.018. Additionally, no distinct separation among the

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support (Computational Biology Program and CAPES-Agropolis Program) and a graduate scholarship to AROC.

Competing Interests: The authors have declared that no competing interests exist.

H. brasiliensis accessions and the other species from Amazonas was observed. A core collection of 99 accessions was identified that captured the maximum genetic diversity. Rubber tree breeders can effectively utilize this core collection for cultivar improvement. Furthermore, such a core collection could provide resources for forming an association panel to evaluate traits with agronomic and commercial importance. Our study generated a molecular database that should facilitate the management of the *Hevea* germplasm and its use for subsequent genetic and genomic breeding.

Introduction

Over the last decade, the number of *ex situ* collections of cultivated plants has greatly increased. This increase has occurred partially in response to the major alterations caused by the development of human settlements in the areas of origin of the related wild species. Among the numerous existing collections, those involving perennial crops (either forestry or fruit crops) are especially difficult to establish. Moreover, these collections are expensive to manage due to the large areas that are required to maintain the accessions, such as the *ex situ* collections of grapevines [1], olive trees [2] and apple trees [3]. These difficulties are often exacerbated in perennial crops from tropical or sub-tropical areas due to the velocity with which ecosystems are deteriorating in these regions and the frequent paucity of the financial resources available for preserving biodiversity in the concerned countries. This issue is of particular concern for cultivated trees originating from Asian tropical rainforests (such as *Tectona grandis* [4]), African tropical rainforests (such as *Elaeis guineensis* [5] and *Coffea canephora* [6]) and Amazonian or Atlantic forests (such as *Theobroma cacao* [7] and *Hevea brasiliensis* [8]). *Ex situ* collections have generally been established to save natural populations of cultivated crops from irremediable destruction when the original habitat for the crop has been threatened by deforestation. These collections also aim to provide breeders with genetic resources that are useful for genetic improvement (e.g., related to resistance to biotic or abiotic stress). However, these *ex situ* collections are frequently partially redundant among countries and even among research centers in the same country due to the duplication of accessions and sharing between institutions [9,10]. Such redundancy is even more common for plants exhibiting easy vegetative propagation [11,12]. The genetic characterization of available genetic resources may permit optimization of the use of these resources by grouping a sufficient number of accessions in a core collection to maximize the genetic diversity described in the whole collection.

The rubber tree (*Hevea brasiliensis*) is a quite recent crop, with genetic improvements initiated only a short time ago. Breeders initiated their work on this species in South East Asia at the beginning of the twentieth century after the successful introduction of approximately 20 seedlings into that part of the world. These seedlings originated from more than 70,000 seeds that were collected by Wickham in 1876 near the Tapajos River in the state of Pará, Brazil. These seeds were sent to England and germinated in the Royal Botanical Garden of Kew, London. The few seedlings obtained after rather poor germination (due to the lack of dormancy in rubber seeds) were sent to Malaysia and Ceylon and served as the origin of all of the rubber tree plantations in the region. The multiplication of these plants was primarily performed through seeds and bud grafting from the 1920s onward when this technique was well managed [13]. Almost all of the clonal varieties that are cultivated today, as well as the genotypes that are still under selection, are less than 10 generations removed from their wild ancestors in the Amazonian forest; this time span represents a very small number of generations for

a tree that is cultivated in very large areas. Furthermore, the vast majority of these clones originated from the so-called Wickham trees that were introduced into Asia at the end of the nineteenth century. Therefore, the genetic basis of the cultivated material is quite narrow, even if most of the cultivars currently exhibit sufficient allelic diversity to allow for substantial genetic improvement.

The cultivated species *H. brasiliensis* originated from the Amazonian forest, where it coexists in various places with some of the ten other species of the genus *Hevea*. All of the *Hevea* species are monoecious and display a preferentially outcrossing habit. Natural interspecific hybridization can occur across the genus and generate gene flow, which has led some authors to consider *Hevea* to be a species complex [14]. Various surveys were performed during the twentieth century with the aim of collecting seeds and budwood from wild rubber trees to increase the genetic pool. One of the first collecting expeditions was conducted under the authority of the Peruvian ministry of agriculture in 1940 in the region of Madre de Dios [15]. In 1974, a joint survey was performed by IRCA (Institut de Recherche sur le Caoutchouc, France) and Embrapa (Empresa Brasileira de Pesquisa Agropecuária, Brazil) in the Brazilian states of Acre and Rondônia [16]. In 1981, the IRRDB (International Rubber Research and Development Board) led an international effort to collect large numbers of rubber tree seeds and budwood from the Brazilian states of Acre, Mato Grosso and Rondônia [17–19]. All of the collected material was dispatched between Brazil, Malaysia and Côte d'Ivoire [20] and then further dispersed in these regions. Later, in an attempt to characterize the diversity with isozymes markers, two natural populations that were close to each other and located near Rio Branco (Acre, Brazil) were sampled and conserved *ex situ* in the Brazilian state of São Paulo [21]. In 1995, a common survey conducted by Embrapa (Brazil) and the RRIM (Rubber Research Institute of Malaysia) was organized to collect seeds from various species of the *Hevea* genus in the Brazilian states of Para and Amazonas [22]. The numerous collected seeds were divided between Malaysia and Brazil. In Brazil, a unique *ex situ* collection containing this material was established in an experimental area from Embrapa near Brasília. To date, the germplasm collected in 1995 has never been described, multiplied or used in breeding programs.

Our objective in the present study was to describe the genetic diversity of the *Hevea* spp. complex available in the main *ex situ* collections of South America. The accessions that were collected during the four surveys described above and conserved in the Brazilian states of Brasília and São Paulo, as well as commercial cultivars and their hybrids were characterized with molecular markers and compared with similar data obtained in a previous study of accessions from an *ex situ* collection in French Guiana. Genetic data were analyzed to describe the genetic structure of the wild populations that were sampled, to quantify the allelic diversity and to suggest the composition of a core collection that would capture the maximum genetic diversity within a minimal sample size.

Materials and Methods

Data from genotypes that were previously collected and analyzed by Le Guen et al. were included in the present work [23]. These data were added to a set of original data obtained from new samples of vegetal material and analyzed jointly. These new data are described hereafter.

Plant material

The numbers of accessions per geographical origin and *ex situ* collection site are listed in Table 1. The majority of these accessions were collected during one of the above-mentioned surveys in Peru (region of Madre de Dios) or Brazil (states of Acre, Amazonas, Mato Grosso,

Table 1. Number of accessions per geographical origin and *ex situ* collection site.

Regions from which seeds were collected	Number of sampled locations	<i>Ex situ</i> collection where the accessions were conserved							Total
		CIRAD Guiana	EMBRAPA CPAC	EMBRAPA Manaus	IAC Pindorama	IAC Votuporanga	Michelin Mato Grosso	Unesp Ilha Solteira	
Acre	6	93	-	-	21	-	12	61	187
Amazonas	7	-	434	-	-	-	-	-	434
Madre de Dios	1	8	-	-	-	-	-	-	8
Mato Grosso	5	117	-	-	2	3	11	-	133
Para	2	-	55	-	-	-	-	-	55
Rondônia	8	129	-	-	20	14	23	-	186
Wickham x Amazonian wild	-	29	-	-	-	-	-	-	29
Wickham	-	48	-	-	7	16	-	-	71
Hevea spp.	-	8	-	6	-	-	-	-	14
Total		432	489	6	50	33	46	61	1117

doi:10.1371/journal.pone.0134607.t001

Pará and Rondônia). Some cultivated clones, all of which were derived from the first introduction of rubber trees to Asia, were also included; these clones were grouped together under the category "Wickham" clones. We separately identified genotypes issuing from crosses between some of the Wickham clones and the accessions of wild Amazonian origin.

The majority of the accessions studied were assumed to be *H. brasiliensis*, with the exception of the following two cases:

- i. A total of 14 accessions from 7 different species of the genus *Hevea* (4 *H. benthamiana*, 2 *H. camargoana*, 1 *H. guianensis*, 4 *H. pauciflora*, 1 *H. spruceana*, 1 *H. paudari* and 1 *H. nitida*) conserved in the *ex situ* collections of Cirad (Guyane) and Embrapa (Manaus). All 14 of these accessions from early surveys have been fully described, and the affiliation of these accessions with the assumed species is not doubted.
- ii. A total of 138 accessions conserved at Embrapa CPAC and collected during the joint Embrapa/RRIM survey of 1995 in Amazonas. These accessions were assumed to originate from *H. brasiliensis* × *H. guianensis* crosses (100 accessions), *H. benthamiana* × *H. brasiliensis* crosses (21 accessions), pure *H. benthamiana* (13 accessions) or pure *H. pauciflora* (4 accessions).

For the majority of the accessions studied, passport data indicating the collection site were available. These accessions were collected from a large number of localities within a large geographical sampling area that is representative of the distribution area of the genus *Hevea*. The collection sites are illustrated in Fig 1.

The detailed data on the accessions are available in S1 Table. Notably, 52 accessions were putatively analyzed twice (i.e., the accessions appeared under the same name but originated from different collection sites). These repeated accessions will provide further insight into the reliability of the applied methods of genetic analysis and the confidence that can be achieved when conserving the name of an accession across various *ex situ* collections.

DNA extraction

DNA extraction was performed using two different methods according to the origin of the samples. For accessions from the Cirad (Guyane) collection, DNA was isolated using a

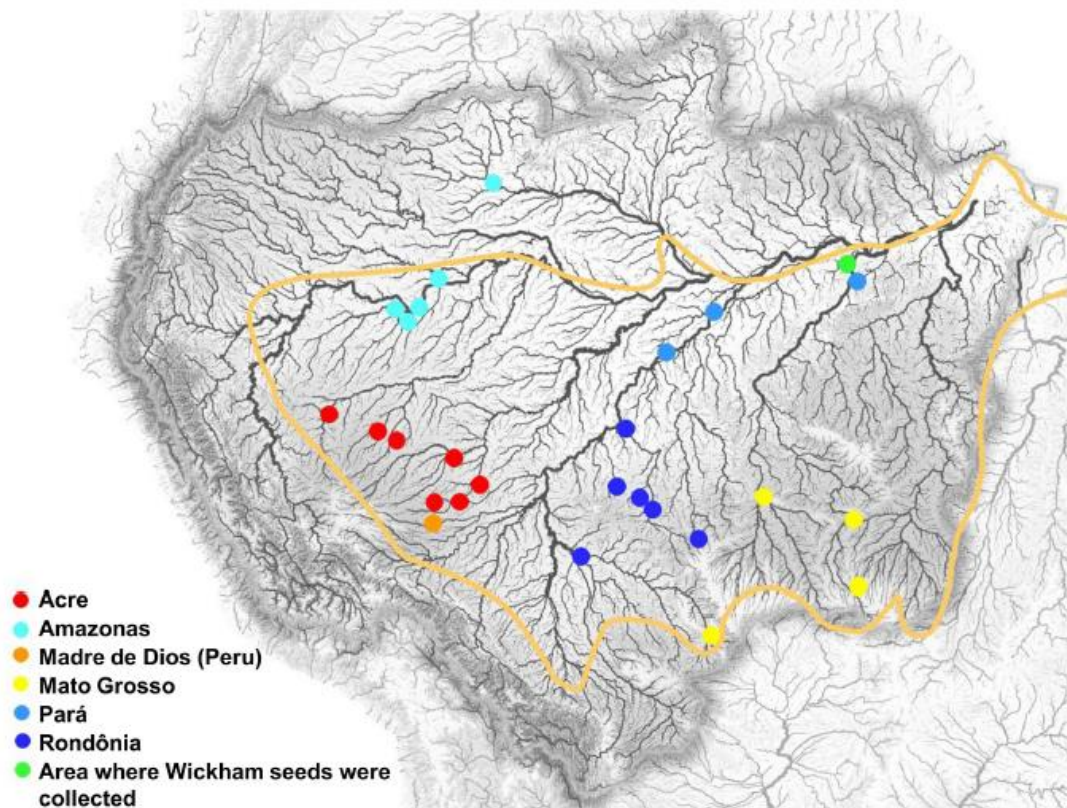


Fig 1. The locations of the sampled sites within Brazil. The colors correspond to the population codes listed in Table 1 and other Figs. This Fig is similar but not identical to the original image from <http://www.mappery.com/map-of/Amazon-Basin-Hydrosheds-Map>.

doi:10.1371/journal.pone.0134607.g001

standard MATAB process as previously described [23]. For the remaining accessions, the collected leaves were stored at -80°C prior to DNA extraction. Genomic DNA was extracted from leaf disks that were 20 mm in diameter and homogenized in a TissueLyser (Qiagen, Valencia, CA, USA). The DNA was extracted according to the Qiagen DNeasy Plant Mini Kit protocol. The quality and concentration of the extracted DNA were analyzed on 1% (m/v) agarose gels.

Analysis of molecular markers

A set of 13 microsatellite (simple sequence repeat—SSR) markers (A2365, A2368, A2387, A2389, A2406, A2413, A2508, A2684, A2736, BAC55B02, T2083, TA2163 and TAS2558) developed in previous studies [23] was selected due to the position of the markers in the *Hevea* genetic linkage map, the large number of alleles and the ease of genotyping.

Polymerase chain reaction (PCR) was performed using a three primer labeling system [24] in a final volume of 10 μl containing 25 ng of template DNA, 0.2 μM each primer, 200 μM

deoxynucleotide (dNTP), 2 mM $MgCl_2$, 1× PCR buffer and 1U of Taq DNA polymerase. The touchdown cycling conditions consisted of an initial denaturing step of 5 min at 94°C, followed by 10 cycles of denaturation, annealing and elongation with a 0.5°C decrease in the annealing per cycle from 55°C to 50°C (94°C for 45 s, 55°C for 1 min, 72°C for 1 min 15 s), 25 cycles with annealing at 50°C (94°C for 45 s, 50°C for 1 min, 72°C for 1 min), and a final elongation step at 72°C for 30 min.

Two different methods were used to separate amplicons based on their migration and visualization. The amplicons from 353 accessions from the CIRAD (Guyane) collection were denatured for 3 min at 94°C, separated on a 6.5% polyacrylamide gel and visualized with a LI-COR 4300 DNA Analyser (LI-COR, Lincoln, NE, USA) as previously described by Le Guen et al. (2009). These 353 accessions included the 307 genotypes analyzed in the above-mentioned study and an additional 46 accessions (mainly from the “Wickham” and “Wickham × Amazonian wild” categories) that, although they were analyzed using the same method and at the same time, were not mentioned in the previous study. The amplicons from the remaining 764 accessions were denatured with formamide and analyzed with an ABI 3500 (Applied Biosystems, Foster City, CA, USA) Automated Sequencer. The alleles were scored against the internal GeneScan-600 (LIZ) Size Standard Kit (Applied Biosystems, Foster City, CA, USA) using GeneMapper software v.4.1.

Characterization of genetic diversity

Analysis of population genetic parameters. A descriptive statistical analysis for all of the SSR loci and the number of alleles per locus (N_a), information index (I), observed heterozygosity (H_o), expected heterozygosity (H_e), fixation index (F) and private alleles (P_a) were calculated using GenAlEx v. 6.5 [25]. The confidence intervals of the different parameters were calculated using 1,000 bootstrap iterations.

The analysis of molecular variance (AMOVA) and calculation of F_{ST} and Nei's G_{ST} among the groups were performed with GenAlEx v. 6.501 [25] to detect the molecular genetic variances within and among the accessions, considering both the collection sites (geographical regions) and the groups suggested by our analysis with the program STRUCTURE. The GenAlEx software was also used to identify private and rare alleles (frequency < 0.05).

Genetic structure analysis. Different statistical procedures were used to describe the genetic diversity of the collected germplasm, including (i) a model-based method for inferring population structure by clustering genotypes; (ii) a distance-based model using a dissimilarity matrix; and (iii) a factorial analysis through principal coordinates analysis (PCoA).

The model-based approach consisted of a Bayesian analysis implemented under the STRUCTURE program version 2.3.4 [26,27]. All 1,117 of the genotyped accessions were analyzed using an admixture model with independent allele frequencies in each group. The number of groups (K) to which the genotypes were assigned was systematically tested from 1 to 10; for each value of K , 20 independent iterations of the Bayesian model-based analysis were performed. The burn-in time and replication numbers were set to 100,000 and 1,000,000, respectively, for each run. We used the ΔK *ad hoc* method described by Evanno et al. (2005) and implemented in the online tool Structure Harvester [28] to estimate the most probable K .

After estimating the most probable K , we utilized the greedy algorithm implemented in CLUMPP v.1.1.1 [29] with a random input order and 1,000 permutations to align the runs. The results were visualized using Distruct v.1.1 [30]. Based on the posterior probability of membership (q) of a given accession belonging to a given group compared with the total (K), we classified individuals showing $q > 0.70$ as members of a given cluster. In contrast, the accession was classified as admixed in clusters with a membership of $q \leq 0.70$. Analyses of

secondary structures were performed (with $K = 1-10$) using only the genotypes corresponding to the clustered group to detect sub-structuring.

The distance-based model was implemented in the DARwin 5.0.158 software [31]. A dissimilarity matrix was computed with a simple-matching index, and a neighbor-joining tree was constructed. A factorial analysis (PCoA) was performed using GenAlEx [25]. These latter methods were used simultaneously to better investigate the precise structure of our sample.

Core collection sampling

The maximization strategy [32] implemented in the software COREFINDER was used to generate a core rubber tree collection that maximized the number of observed alleles in our dataset. The M-strategy consisted of detecting the sample size that best captured 100% of the genetic diversity present within the entire germplasm collection.

Ethics statement

We confirm that no specific permits were required to collect the leaves used in this study. This work was a collaborative study developed by researchers from APTA (Brazil), CIRAD (France), EMBRAPA (Brazil), IAC (Brazil), UESB (Brazil) UNESP (Brazil) and UNICAMP (Brazil). Additionally, we confirm that the field studies did not involve endangered or protected species.

Results

Data homogenization

As explained in the previous section, the allelic sizes of the PCR products were estimated using the following two technologies: LI-COR technology (electrophoresis in polyacrylamide gels) for 353 accessions and ABI3500 (capillary electrophoresis) for the remaining 764 accessions. Although both technologies are accurate and reliable, deviations may exist between the data obtained from these two sources due to the use of different calibration curves. The analysis of common control genotypes using both technologies allowed us to standardize the sizes of the different allelic products for each microsatellite marker. Moreover, the presence of 52 pairs of accessions among the analyzed samples identified under the same name and analyzed with the two technologies helped us to confirm the goodness of data homogeneity. Among these 52 pairs of homonymic accessions, 26 were identical, 22 exhibited less than 5 differences among the 26 allelic forms (13 microsatellites with 2 alleles each), and 4 differed for at least 5 alleles. Further analyses based on a matrix of dissimilarity (see below) suggested that only these last 4 pairs corresponded to genetically distinct genotypes to which identical names were erroneously assigned.

Genetic diversity

We computed diversity statistics across all 1,117 of the individual accessions and within each of the two subgroups obtained in the analyses using the STRUCTURE software (Table 1). The overall number of different alleles per SSR locus across all of the accessions ranged from 5 to 39 (data not shown), with a group mean of 14.5 (Table 2).

For all of the accessions, the mean observed heterozygosity (H_O) was 0.64 (ranging from 0.57 to 0.69); this value was slightly higher in cluster 1 than in cluster 2. The genetic diversity (H_E) was higher than H_O in all cases. The analysis revealed a high H_E level ranging from 0.63 (Wickham group) to 0.84 (Pará group) with a mean of 0.76. Consequently, the Wright's fixation index (F) values were positive, with a mean of 0.16 obtained for the accessions overall. The

Table 2. Descriptive results of the values observed in the characterization of the mean loci for each population of sample size (N), number of alleles (Na), information index (I), observed heterozygosity (H_o), expected genetic diversity (H_E), fixation index (F) and private alleles (Pa) in 1,117 accessions of *Hevea* spp.

Groups ^a	N	Na	I	H_o	H_E	F	Pa
Acre (AC)	183	20.0	2.24	0.66	0.81	0.18	11
Amazonas (AM)	428	23.8	2.29	0.65	0.82	0.21	34
Wickham (Wi)	70	9.5	1.39	0.57	0.63	0.09	0
Madre de Dios (MD)	7	6.4	1.53	0.67	0.70	0.05	0
Mato Grosso (MT)	129	14.3	1.88	0.59	0.74	0.23	7
Para (PA)	54	15.8	2.23	0.69	0.84	0.17	3
Rorônia (RO)	183	21.1	2.27	0.63	0.81	0.22	14
Wickham x Amazonian wild (Wi x AM)	28	9.5	1.61	0.64	0.68	0.08	0
<i>Hevea</i> spp.	13	10.2	2.05	0.67	0.83	0.18	20
Mean	122	14.5	1.94	0.64	0.76	0.16	10
SE	11.7	0.74	0.06	0.02	0.02	0.01	11.5
Cluster ^b 1 (red)	24	8.88	2.36	0.67	0.84	0.20	18
Cluster 2 (green)	30	9.04	2.41	0.63	0.84	0.25	97

^a The 1,117 accessions were divided according to the locations of origin from which they were collected and assigned to groups.

^b Structure analysis results suggest that the set of genotypes can be partitioned into two clusters.

doi:10.1371/journal.pone.0134607.t002

heterozygosities (H_o and H_E) were similar in both of the subgroups, although cluster 2 displayed higher indices of fixation ($F = 0.20$ in cluster 1 and $F = 0.25$ in cluster 2).

Of the 408 observed alleles, 319 alleles were shared among the groups; the remaining 89 alleles represented private alleles in different groups of accessions (Fig 2). The A2508 locus contained the largest number of private alleles of the 13 SSR alleles used in this study (15). The

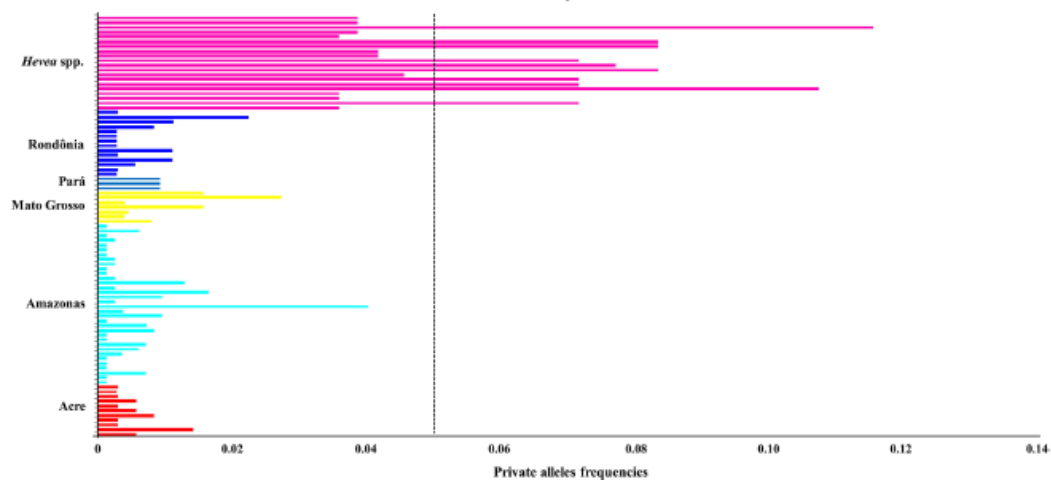


Fig 2. Private allele frequencies obtained from the genotyping of microsatellite loci in the *Hevea* germplasm. The dashed line indicates the cutoff for the occurrence of rare alleles (frequency = 0.05).

doi:10.1371/journal.pone.0134607.g002

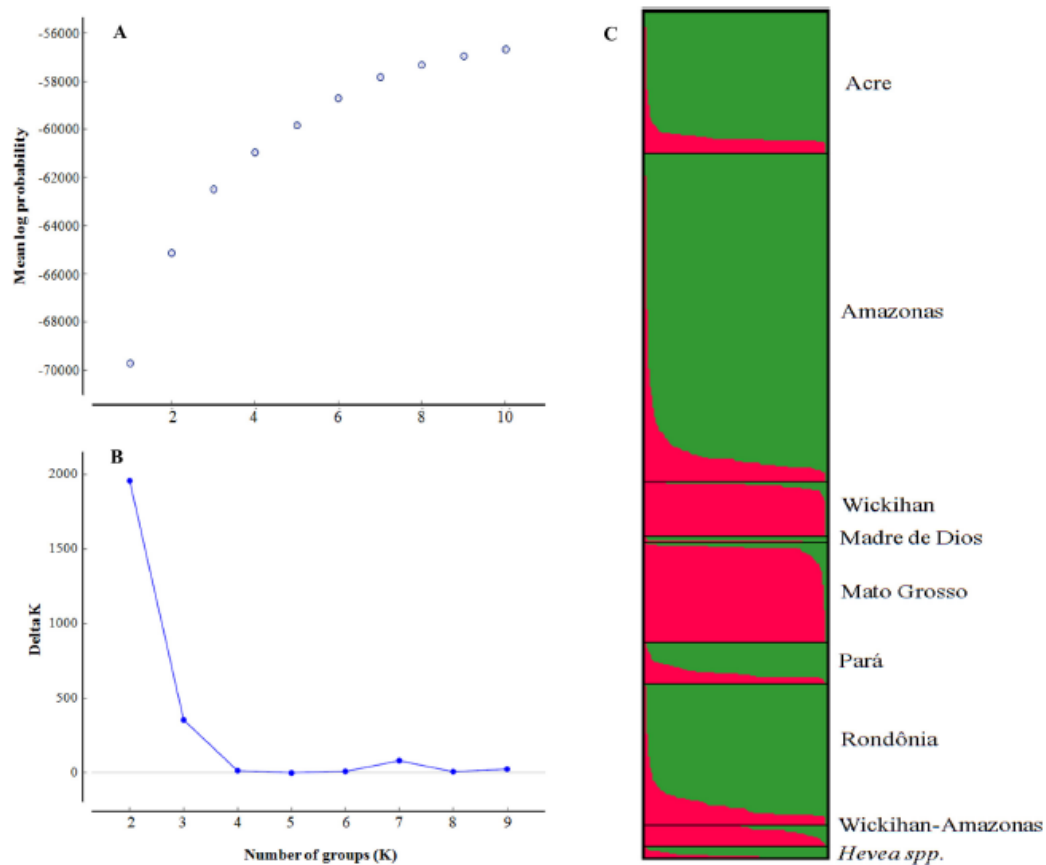


Fig 3. Genetic structure of the *Hevea* germplasm accessions. The genetic structure was inferred through Bayesian analyses of the most probable number of groups (K) estimated based on the mean log probability [A] using the method described by Evanno et al. (2005) [B]. Each column (histogram) [C] represents the genotyping data from one accession, and the colors used in the histogram represent the most likely ancestry of the cluster from which the accessions were derived.

doi:10.1371/journal.pone.0134607.g003

Amazonas population and *Hevea* spp. population exhibited the greatest number of private alleles (34 and 20, respectively). A comparison of the number of private alleles from cluster 1 (18) and cluster 2 (97) showed that the latter contained a larger number of private alleles (Table 2, Fig 2).

Model-based investigation of genetic structure

To examine the relatedness among the 1,117 accessions, the data for 13 markers were analyzed using the STRUCTURE program. A strict interpretation of our results using the method of Evanno et al. (2005) suggested that two genetic clusters ($K = 2$) were sufficient to interpret our data (Fig 3A and 3B); the results are also presented as $K = 3$ and $K = 7$ in S3 Fig. Based on a

Table 3. Division of membership probability in the two clusters (K = 2) inferred based on Bayesian analyses considering the most probable number of groups (K); designated as Cluster 1, Cluster 2 and Mixture (membership $q < 0.07$).

Groups ^a	Total	Cluster 1 (red)	Cluster 2 (green)	Mixture
Acre (AC)	187	17	165	5
Amazonas (AM)	434	22	400	12
Wickham (Wi)	71	67	2	2
Madre de Dios (MD)	8	1	7	0
Mato Grosso (MT)	133	124	6	3
Para (PA)	55	9	39	7
Rondônia (RO)	186	15	154	17
Wickham x Amazonian wild (Wi x AM)	29	22	0	7
<i>Hevea</i> spp.	14	0	10	4

^a The 1,117 accessions were divided according to the locations of origin from which they were collected and assigned to a group.

doi:10.1371/journal.pone.0134607.t003

membership probability threshold of 0.70, 277 accessions were assigned to cluster 1 and 783 accessions were assigned to cluster 2. The remaining 57 accessions were classified into a mixed subgroup. The membership probability was < 0.70 in any given subgroup (Table 3).

The level of clustering (K = 2) reflects the primary division of subgroup cluster 1, consisting of the varieties of the advanced breeding germplasm originating from the Wickham and Mato Grosso accessions. Cluster 2 was formed by the wild germplasm from the Acre, Amazonas, Pará and Rondônia populations and *Hevea* spp. A further STRUCTURE analysis of these two primary clusters of genotypes also indicated partitioning of both cluster 1 and cluster 2 into 2 subgroups (Fig 3). However, no clear pattern for the assignment of individuals based on geographical origin was observed.

Additionally, the separation of the groups was also estimated using PCoA (Fig 4A) based on the genetic distance, which divided the entire population into two clear clusters related to the structure analyses. The first and second principal coordinates explained 35% and 20% of the molecular variation, respectively (Fig 4A and 4B). The accessions from Acre, Amazonas, Madre de Dios, Pará, Rondônia and *Hevea* spp. were placed into one group, while the remaining populations belonged to the group formed by the Wi, MT and Wi x AM wild accessions.

Distance-based analysis

The dissimilarity matrix computed with DARwin 5.0.158 was used to identify putative unexpected duplications of genotypes under different accession names. A total of 13 pairs and 2 triplets of accessions were found to be present with a high degree of genetic similarity. All of the concerned accessions were from the oldest and most widespread surveys conducted in 1974 (by EMBRAPA/IRCA) and 1981 (by IRRDB). In the majority of cases, these accessions were registered under almost identical names (e.g., MT/C/5/25 vs. MT/C/S/25 or RO/C/9/45 vs. RO/C/9/48). Therefore, misidentifications due to human mistakes can be reasonably hypothesized to have occurred during the process of multiplication or transfer of accessions from one *ex situ* collection to another. In two cases, the accession names were completely different.

Finally, we constructed a neighbor-joining tree to confirm the relationships among the accessions. The tree (Fig 5A) showed a pattern that was consistent with the two analyses described above. The group formed by the Wi, MT and Wi x AM wild accessions was clearly separated. The two subgroups suggested by the STRUCTURE analyses were also indicated in the tree. A discrete separation was observed between the genotypes of *H. brasiliensis* and other

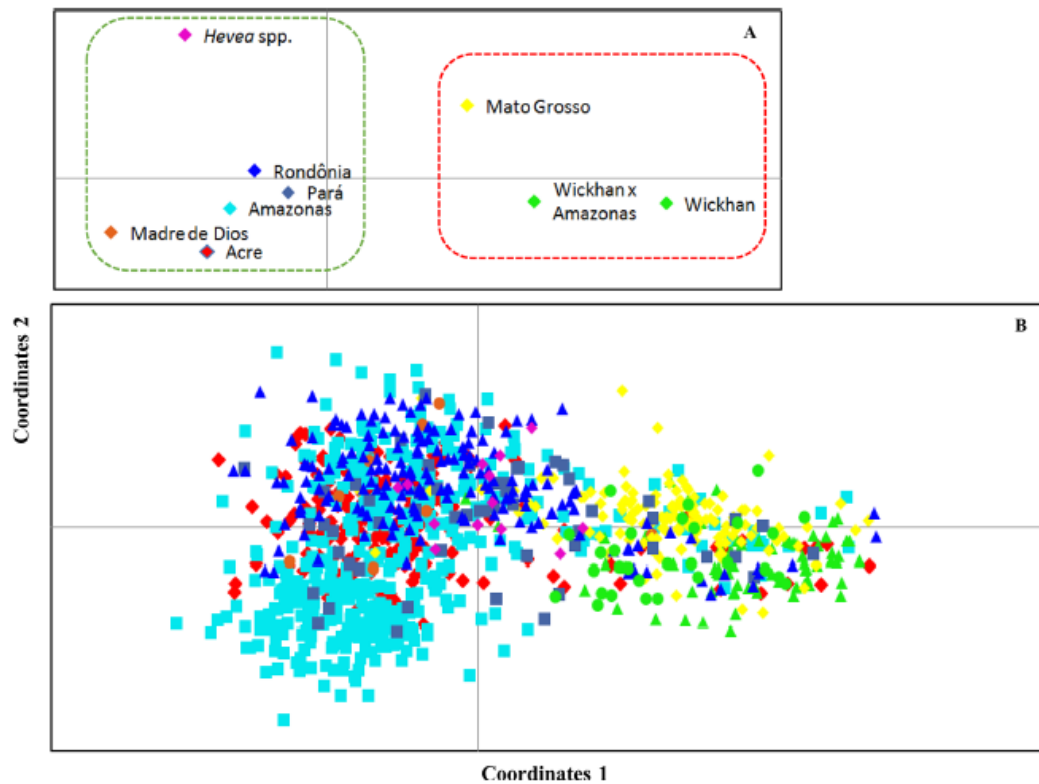


Fig 4. Dispersion graphics based on a principal coordinates analysis (PCoA) of the groups. [A] Accessions (the dashed lines refer to the groups that were observed based on the analysis conducted using STRUCTURE software) [B] from the *Hevea* germplasm based on genotypic data from 13 microsatellite loci.

doi:10.1371/journal.pone.0134607.g004

species or between the other species of the genus *Hevea* (13 *H. benthamiana* and 4 *H. pauciflora*). However, the hybrids (21 *H. brasiliensis* × *H. benthamiana* and 100 *H. brasiliensis* × *H. guianensis*) were not separated from the genotypes of the Amazonas group (Fig 5B); these results were confirmed by the PCoA (Fig 4).

We used AMOVA to partition the molecular genetic variance within and among the accessions and between the nine subgroups according to their country of origin. In a preliminary AMOVA, 7% of the total genetic variance was found to occur between the groups, while 21% occurred between the accessions and 73% occurred within the accessions.

Although most of the genetic variance occurred within the accessions, the differentiation of the accessions and groups was highly significant ($p \leq 0.001$). The pairwise G_{ST} values (analogous to F_{ST}) ranged from 0.008 (between the Amazonas and PA) to 0.099 (between the Wi and MD). Each of the pairwise values was significantly different from zero according to tests based on 9,999 random permutations ($p \leq 0.001$) (Table 4). This finding was similar to the F_{ST} values, which ranged from 0.015 (between the AM and PA) to 0.20 (between the Wi and MD).

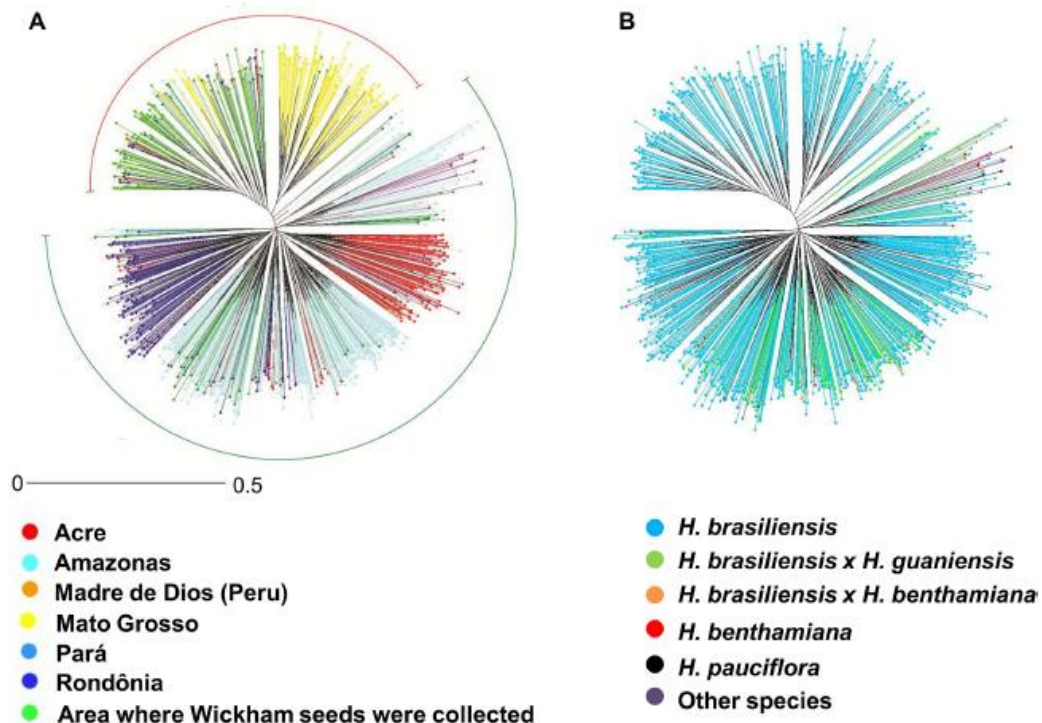


Fig 5. Neighbor-joining tree as a third method of visualizing the relationships between the genotypes. [A]. Classified based on the region of collection. [B]. Classified by the species group.

doi:10.1371/journal.pone.0134607.g005

However, the sample sizes from MD are relatively small compared to the other accessions, which may influence the F_{ST} values. In terms of the population structure between the groups

Table 4. Triangular matrices for the computation of G_{ST} and F_{ST} tests among the *Hevea* germplasm; G_{ST} values are below the diagonal, and F_{ST} values are above the diagonal.

Groups	AC	AM	Wi	MD	MT	PA	RO	WxAM	<i>Hevea</i> sp.
Acre (AC)	—	0.028	0.138	0.036	0.095	0.030	0.044	0.085	0.083
Amazonas (AM)	0.015	—	0.131	0.064	0.089	0.015	0.039	0.088	0.071
Wickham (Wi)	0.081	0.079	—	0.201	0.086	0.100	0.134	0.036	0.179
Madre de Dios (MD)	0.010	0.025	0.099	—	0.132	0.061	0.076	0.117	0.086
Mato Grosso (MT)	0.053	0.049	0.047	0.063	—	0.061	0.076	0.056	0.109
Pará (PA)	0.016	0.008	0.054	0.027	0.032	—	0.033	0.058	0.065
Rondônia (RO)	0.023	0.020	0.078	0.031	0.041	0.017	—	0.089	0.076
Wickham x Amazonian wild (WxAM)	0.049	0.051	0.018	0.057	0.031	0.032	0.050	—	0.124
<i>Hevea</i> sp.	0.043	0.036	0.091	0.052	0.055	0.034	0.038	0.067	—

doi:10.1371/journal.pone.0134607.t004

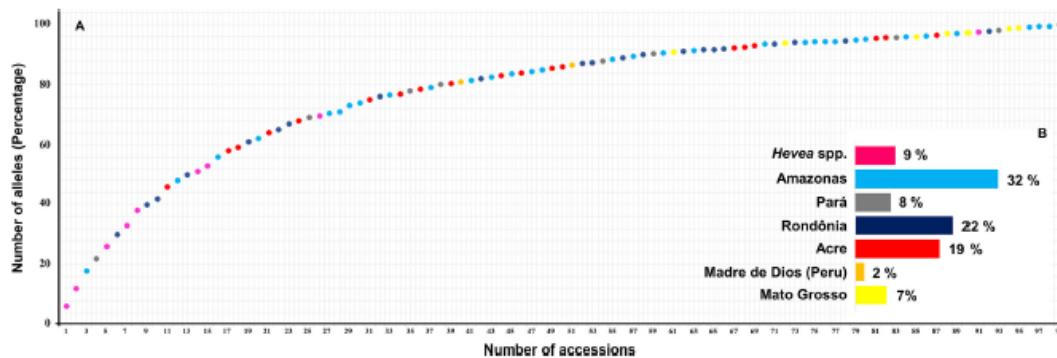


Fig 6. Genetic diversity (expressed as a percentage of the total number of alleles) as a function of the number of accessions included in the rubber tree germplasm core collection.

doi:10.1371/journal.pone.0134607.g006

recognized in this study, the G_{ST} between cluster 1 and cluster 2 estimated from AMOVA was 0.018 ($P < 0.01$).

Rubber tree core collection

A core collection was assembled for the rubber tree germplasm to represent all of the genetic diversity identified in this study. The 408 alleles identified at the 13 microsatellites used in this study were fully represented by as few as 99 accessions (the best of 10 runs) (Fig 6 and S1 Table). The core collection was represented by the accessions from Amazonas (32.3%), Rondônia (22.2%), Acre (19.2%), *Hevea* spp. (9.1%), Pará (8.1%), Mato Grosso (7.1%) and Madre de Dios (2%). The alleles identified in this study were fully represented in these core collections. Additionally, our COREFINDER analysis highlighted that 85% (84 accessions) of the entire core collection was represented by the rubber tree accessions grouped in cluster 2, whereas cluster 1 contributed a smaller percentage (11%) to the core collection; the accessions that were considered a mixture (classified into this subgroup by Structure analysis) contributed to only 4% (4 accessions) of the core collection.

Discussion

The history of the collections

The present study is the most comprehensive study ever performed on the population genetics of the genus *Hevea*. This study was based on more than 1,100 accessions sampled at 7 *ex situ* collection sites located in Brazil and French Guiana. Nearly 500 accessions from the collecting survey conducted in 1995 in the Brazilian states of Amazonas and Pará that had never been previously described were included in this study. The other accessions mostly came from the states of Acre, Mato Grosso or Rondônia and were obtained in surveys performed in 1974 and 1981. Since that time, the accessions have been conserved in different *ex situ* collections. In recent decades, this latter group of accessions has potentially been transferred from one collection site to another, with all of the possible associated risks regarding traceability. For example, the accessions kept at the French Guiana collection site were primarily contributed by the CNRA collection site at Bimbresso in Côte d'Ivoire, where they had been sent soon after the respective prospection [20]. Some of these accessions were subsequently distributed from

French Guiana to the Michelin collection in Brazil. Therefore, the first use of highly polymorphic microsatellite markers to characterize this material was to determine the overall consistency of the accession denominations. The comparison of 52 pairs of homonymic accessions conserved at different sites allowed us to conclude that only 4 corresponded to distinct genotypes. Moreover, 13 pairs (and 2 triplets) of accessions with differing identifications presented highly similar genotypes, suggesting loss or modification of their true identity across the transfer of accessions between collection sites. Notably, these anomalies in identification were detected only for old accessions that were transferred from site to site through grafted plants. The more recent accessions (i.e., from Amazonas and Pará in EMBRAPA/CPAC or from Acre in UNESP Ilha Solteira) that have not yet been multiplied did not present these anomalies. Therefore, microsatellite markers appear to be useful tools for the management of accessions in collections, especially when transfers or exchanges of biological material between sites are scheduled.

Genetic diversity

The results from this study reveal high levels of genetic diversity among the accessions, which is consistent with previous studies [23,33] and the biology of this outbred plant species. We detected an average of 14.5 alleles per locus and an H_o of 0.64 across the entire accession set. These values were slightly higher than those detected by Le Guen et al. (2009) ($H_o = 0.59$). Lam et al. (2009) have detected an average heterozygosity of 0.20 using random amplified polymorphic DNA (RAPD) markers in *Hevea* genetic resources in Vietnam. These differences in genetic diversity values might be attributable to the types of germplasm that were analyzed. Lekawipat et al. (2003) have used microsatellite markers to detect the diversity in 40 Wickham and 68 Amazonian accessions using 170 alleles from 12 microsatellite markers dispersed among all of the genotypes. Similar results were observed in our study, in which an average of 14 alleles were available per locus; this diversity can be attributed to the high variability of Amazonian accessions.

The estimates of G_{ST} and F_{ST} values revealed low genetic differentiation between the groups, with the exception of the Wickham accessions, which showed greater differentiation compared to the other studied groups (Table 4). The Wickham accessions originated from breeding programs with a narrow genetic base [8,34]; thus, the Wickham accessions were unambiguously less variable than the Amazonian accessions, thus causing the inflation of F_{ST} .

An outcrossing breeding system, long life cycle and large geographic range may have helped to shape the high genetic diversity of *Hevea* plants. Le Guen et al. (2009) have obtained a similar value ($F_{ST} = 0.12$) upon analyzing *ex situ* collections of wild rubber trees. Gouvea et al. (2010) have studied clones from controlled-cross and open pollinated progenies and reported high values of genetic differentiation ($G_{ST} = 0.61$). Lower genetic differentiation indicates higher gene flow between populations, indicating that a majority of the genetic diversity is preserved within populations.

A population structure analysis using the software package STRUCTURE revealed the presence of two primary groups in our set of accessions. This result partially agrees with the results obtained through PCoA (Fig 4) analysis, in which the genotypes from *Hevea spp.* were discretely separated from the other populations to form cluster 2. However, due to the small number of samples analyzed (only seven *Hevea spp.* genotypes were included in this group), this result should be interpreted with caution. The observed heterozygosity was slightly higher in cluster 1 (Wi, Wi × AM and MT accessions) than in cluster 2 (formed by the other analyzed accessions); nevertheless, the number of different alleles was greater in cluster 2, which can be explained by the greater diversity of the different species present in this cluster.

Prior analyses by Le Guen et al. (2009) have demonstrated a distinct separation between the Acre and Rondônia groups and the Mato Grosso group. Although only a few Wickham accessions were included among the samples, these clones appeared to be closer to the Mato Grosso accessions. Our analysis of 55 additional Asian accessions confirmed the genetic similarity between the Wickham and Mato Grosso accessions. The most obvious division in the analyzed accessions occurred between the Mato Grosso and Wickham clones; these clones formed one group, while the other accessions grouped together in a second cluster.

One explanation for this result is that the accessions are from geographically distant populations. However, these two groups are located in the same hydrographical network, which originates downstream of the Tapajós River [23,34,35].

Determining a structure within the accessions of Amazonas, Rondônia, Acre, Madre de Dios and Pará is more difficult. One possible explanation for this difficulty is that long-distance gene flow via seed flotation prevented the isolation of Amazonian populations of *H. brasiliensis* separated by distances within catchments and incurred the isolation in populations with distances that exceeded the size of a catchment [23]. The rubber tree is pollinated by midges and thrips, and *Hevea* pollen can generally be dispersed by these pollinating agents in the course of open pollination. Yeang et al. (1999) [36] have estimated that the proportion of crosses and selfs changes with distance, and *Hevea* pollen can be transported fairly considerable distances by pollinating agents. The regression model in this study suggested a theoretical range of 0 ± 3 to 1 ± 1 km, although very few pollen grains would be expected to reach this limit.

The first study aiming to understand the genetic diversity of the wild *Hevea* germplasm was conducted using isozyme markers in accessions from the 1981 IRRDB collection and Wickham group. The results of the isozyme analysis detected a total of 60 alleles in the 1981 IRRDB accessions and 26 alleles in the Wickham accessions. Based on these results, Chevallier (1988) affirmed that the genetic base of the *Hevea* germplasm 1981 IRRDB collection was prosperous and diversified. In contrast, the Wickham population showed a low level of genetic variability, which is a consequence of oriented selection within a narrow geographical origin over a period of years. Our results demonstrated a greater number of private alleles in cluster 2, which is composed of the wild germplasm (Table 2, Fig 2), thus confirming that the wild germplasm accessions bear most of the rubber tree genetic diversity present in the collections.

Other studies have subsequently been performed on the same 1981 IRRDB population using restriction fragment length polymorphism (RFLP) markers [35,37] and PCR-derived markers [23,34]. This collection can contribute effectively to the long-term progress of the *Hevea* breeding and selection program due to its high genetic variability.

In 1995, an expedition was launched by RRIM to collect rubber seeds from Brazil. From this collection, approximately 50,231 seedlings were planted in Malaysia, including other species of the *Hevea* genus. The first study using this material selected 150 genotypes to identify the most promising rubberwood-yielding genotypes and wild *Hevea* relatives conserved in the 1995 *Hevea* germplasm in Selangor, Malaysia. The genotypes and wild relatives that are considered vigorous and generate a high rubberwood yield can be exploited as promising planting material in breeding programs. In short, the expedition to collect wild *Hevea* relatives in 1995 was a worthwhile scientific activity to safeguard the most valuable genetic resources and improve future crops of the *Hevea* species [22].

Although this *ex situ* germplasm collection presents great genetic potential for breeding, the collection remains underutilized. AMOVA demonstrated that 3% of the total genetic variance was present between the groups and 97% was within the groups, which could be explained by the large number of analyzed individuals in each group. Although the majority of the genetic variation occurred between the individuals within a group, the differences between the accessions and group differentiation were significant ($p < 0.001$). Our analysis revealed a high

frequency of gene flow between the groups, with a genetic differentiation coefficient (G_{ST}) estimated to be 0.018.

The majority of the germplasm collection used in the current breeding programs was derived from the 1981 IRRDB collection expedition in the Amazon forests of Brazil, but 434 clones from the 1995 germplasm collection (Amazonas genotypes) showed 34 private alleles and high genetic variability, which stands out among the groups. Altogether, the groups contained 55 private alleles, of which 20 alleles were from the *Hevea* spp. group. These genotypes may be incorporated in future breeding programs.

A species complex

In addition to the high gene flow detected, no distinct separation among the *H. brasiliensis* accessions and the other species from Amazonas was observed in the Structure analysis. However, using PCoA analysis, a possible discrete separation of *Hevea* spp. genotypes from the other genotypes of cluster 2 was observed. The results obtained by Lam et al. (2009) suggest that gene flow might exist between the districts, possibly due to the species' outcrossing mode of reproduction and the dispersion of seeds by a network of rivers in the Amazon basin. However, this explanation should be treated with caution due to the small number of samples analyzed in that study.

In the present work, several Amazonas accessions were originally identified as other *Hevea* species (13 *H. benthamiana* and 4 *H. pauciflora*) or interspecific hybrids (21 *H. benthamiana* × *H. brasiliensis* and 100 *H. brasiliensis* × *H. guianensis*). This previous identification was based on the seed size, shape and layout pattern of the accessions. However, whether the trees from which the seeds were collected were hybrids or pure species pollinated by other species is not clear. No methods exist to easily determine whether an *H. brasiliensis* seed was generated by pollination from *H. brasiliensis* or *H. guianensis*. Our difficulty in structuring the Amazonas populations, even within the *H. brasiliensis* genotypes, mostly likely occurs because we are not observing a single species but a species complex consisting of at least *H. brasiliensis*, *H. guianensis* and *H. benthamiana*.

Since the publication of the first papers on the subject, the *Hevea* species have been generally assumed to be freely interfertile [38,39]. Pires (1981) has observed natural hybrids of *H. camargoana* × *H. brasiliensis*, and Gonçalves et al. (1982) have analyzed the progenies that issued from hand pollination through this type of crossing. Additionally, *Hevea* species might be considered a species complex due to the absence of a strict barrier to recombination between the species.

Recently, studies regarding the cross-species/genera transferability of SSR markers have been reported [40–43]. According to Saha et al. (2005), the cross-fertility potential and high SSR transferability observed in *Hevea* species indicate that the genus is a species complex with moderate differentiation between species. Moreover, the results suggest that these aspects are favorable for genetic introgressions using other *Hevea* species in the rubber breeding population, which is mainly based on *H. brasiliensis*.

A core collection to capture allelic diversity

The conventional solution for the conservation of *Hevea* genetic resources has been the establishment of *ex situ* gene banks that can represent a last resort for many species and varieties (including those of the *Hevea* genus), which would otherwise die out as their habitats are destroyed [33]. The conversion of tropical rain forests to pasture and cropland is having dramatic effects on the environment. Over the past several years, the native landscape in the Amazon basin has been rapidly and radically transformed by human actions. Particularly intense

and rapid deforestation is taking place in the state of Rondônia, Brazil, that can be observed in a series of satellite images (S1 and S2 Figs).

In this work, 186 *Hevea spp.* accessions were collected from 8 Rondônia regions that have most likely been deforested during the last 30 years, resulting in a huge loss of the important *Hevea spp.* germplasm. Considering the deforestation scenario in Amazonia (and particularly in Rondônia), the establishment and *ex situ* conservation of *Hevea spp.* germplasm collections is the last possibility to preserve their genetic diversity.

The purpose of core collections is to facilitate the use, maintenance and management of the germplasm by providing a set of accessions that display most of the genetic diversity available in the larger collection with minimum redundancy [44]. Such strategies have already been developed for important perennial crops such as apple trees [45], olive trees [2] and fig trees [46]; these collections are essential for the rational utilization of the germplasm scattered among numerous *ex situ* collection sites in different countries.

Various *Hevea* gene banks exist worldwide, mostly on the South American continent. However, the majority of these collections are composed of accessions from the 1981 IRRDB survey [47]. None of these accessions contain as much diversity as described in the present study for the accessions collected during the last extensive survey performed in 1995 in the Brazilian states of Pará and Amazonas. The present study addressed this topic by providing a molecular database that should facilitate the management of the *Hevea* germplasm. Moreover, we identified a large number of private alleles and high genetic variability in the accessions from the 1995 germplasm. Although the planted and recently domesticated populations exhibited greater genetic diversity, the wild populations have preserved the most private and rare alleles, causing these populations to be the most important reservoirs of genetic variation.

The selection procedure for developing a *Hevea* core collection resulted in the selection of 99 accessions from the germplasm collection with 85% of the entire core collection represented by the rubber tree accessions grouped in cluster 2. The observed heterozygosity was slightly lower in cluster 2 than in cluster 1, but the number of private alleles was greater in cluster 2, as predicted. These observations can be explained by the greater diversity of the different species present in cluster 2 and the greater contribution to the core collection. A total of 32.3% of the accessions were from the Amazonas group, indicating the importance of the wild *Hevea* relatives collected in 1995 and confirming that these collection efforts were a worthwhile scientific activity to safeguard the most valuable genetic resources for future crop improvements of the *Hevea* species.

The accessions of the proposed core collection comprise 85% wild material, and the collection would possess abundant genetic diversity and private alleles, an advantageous attribute for dissecting the genetic basis of QTL for immediate application in *Hevea* breeding.

These investigations should include complementary criteria, such as phenotypic, agronomic and adaptive traits. A core collection would also provide a logical subset of germplasm for examination when the entire collection cannot be used.

Finally, such core collections would also be useful for the development of new breeding strategies for adaptive and agronomic traits through genome-wide association studies.

Supporting Information

S1 Fig. Satellite images of the state of Rondônia in western Brazil. (A) June, 1975; (B) July, 1986; (C) August, 2001; (D) August, 2013; red dot—city of Ariquemes. As much as 67,764 km² of rain forest had been cleared through 2003. Systematic cutting of forest vegetation starts along the roads and then fans out to create the "fishbone" pattern that begins to be visible in the eastern half of the 1986 image. The deforested land and urban areas appear lavender; healthy

vegetation appears green. (<http://earthshots.usgs.gov/earthshots/node/39#ad-image-0>). Images courtesy of the Geological Survey (the USGS home page is: <http://www.usgs.gov>) (TIFF)

S2 Fig. Deforestation branching off from roads in the city of Ariquemes, state of Rondônia, in western Brazil. (A) June, 1975; (B) July 1986; (C) August, 2001; (D) August, 2013. The early images show main roads cutting through the forest. Highway 421 snakes through the forest south-southwest from the city of Ariquemes, and Highway 364 runs roughly north to south through Ariquemes. Additional roads branch out from the main roads to create the fishbone pattern. As time proceeds, a patchwork of cleared areas, forest remnants, and settlements are left behind. (<http://earthshots.usgs.gov/earthshots/node/39#ad-image-5>). Images courtesy of the Geological Survey (the USGS home page is: <http://www.usgs.gov>). (TIFF)

S3 Fig. Bayesian clustering results. Bar plots from the CLUMPP results aligning 20 structure runs for $K = 3$ and $K = 7$. (TIFF)

S1 Table. Population sampled in this study with the state origin, putative geographic origin, number of accessions sampled and genotyping results obtained with the 13 microsatellite markers. (XLSX)

Author Contributions

Conceived and designed the experiments: LMS VLG PSG APS. Performed the experiments: LMS VLG AROC CCM CCS EJSJ JFF MLTM PSG. Analyzed the data: LMS VLG CBMCS JPGV MIZ. Contributed reagents/materials/analysis tools: APS VLG PSG. Wrote the paper: LMS VLG APS.

References

1. Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon A-F, et al. Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. *sativa*. *BMC Plant Biol*. 2008; 8: 31. doi: [10.1186/1471-2229-8-31](https://doi.org/10.1186/1471-2229-8-31) PMID: [18384667](https://pubmed.ncbi.nlm.nih.gov/18384667/)
2. Belaj A, Dominguez-Garcia M del C, Atienza SG, Martín Urdiroz N, de la Rosa R, Satovic Z, et al. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DARs, SSRs, SNPs) and agronomic traits. *Tree Genet Genomes*. 2012; 8: 365–378.
3. Van Treuren R, Kemp H, Emsting G, Jongejans B, Houtman H, Visser L. Microsatellite genotyping of apple (*Malus x domestica* Borkh.) genetic resources in the Netherlands: Application in collection management and variety identification. *Genet Resour Crop Evol*. 2010; 57: 853–865.
4. Graudal LOV, Kjaer ED, Suangtho V, Saardavut P, Kaosa-ard A. Conservation of genetic resources of teak (*Tectona grandis*) in Thailand. Danida For Seed Cent. 1999; Technical note n° 52.
5. Cochard B, Adon B, Rekima S, Billotte N, De Chenon RD, Koutou A, et al. Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. *Tree Genet Genomes*. 2009; 5: 493–504.
6. Cubry P, de Bellis F, Pot D, Musoli P, Leroy T. Global analysis of *Coffea canephora* Pierre ex Froehner (*Rubiaceae*) from the Guineo-Congolese region reveals impacts from climatic refuges and migration effects. *Genet Resour Crop Evol*. 2013; 60: 483–501.
7. Marciano M, Pugh T, Cros E, Morales S, Portillo Pérez EA, Courtois B, et al. Adding value to cocoa (*Theobroma cacao* L.) germplasm information with domestication history and admixture mapping. *Theor Appl Genet*. 2007; 114: 877–884. PMID: [17252253](https://pubmed.ncbi.nlm.nih.gov/17252253/)
8. Pires JM, Secco RS, Gomes JI. Taxonomia e fitogeografia das seringueiras (*Hevea* spp.). Belém: Embrapa Amazônia Oriental; 2002.

9. Greene S, Pederson G. Eliminating Duplicates in Germplasm Collections: A White Clover Example. *Crop Sci.* 1996; 36: 1398–1400.
10. Shan F, Clarke HJ, Yan G, Plummer JA, Siddique KHM. Identification of duplicates and fingerprinting of primary and secondary wild annual *Cicer* gene pools using AFLP markers. *Genet Resour Crop Evol.* 2007; 54: 519–527.
11. Gross B, Volk G, Richards C, Forsline P, Fazio G, Chao C. Identification of "Duplicate" Accessions within the USDA-ARS National Plant Germplasm System *Malus* Collection. *J Am Soc Hortic Sci.* 2012; 137: 333–342.
12. Zhang D, Mischke S, Johnson ES, Phillips-Mora W, Meinhardt L. Molecular characterization of an international cacao collection using microsatellite markers. *Tree Genet Genomes.* 2009; 5: 1–10. doi: [10.1007/s11295-008-0163-z](https://doi.org/10.1007/s11295-008-0163-z)
13. Van Heltten VM. Het oculeeren van *Hevea*. *Arch voor Rubbercult Ned.* 1918; 2: 87.
14. Priyadarshan PM, Goncalves PS. *Hevea* gene pool for breeding. *Genet Resour Crop Evol.* 2003; 50: 101–114.
15. Rands RD, Polhamus LG. Progress report on the cooperative *Hevea* rubber development in Latin America. U.S.D.A., Washington DC. 1955; 79 p.
16. Hallé F, Combe J-C. en Amazonie brésilienne pour la récolte de matériel génétique nouveau destiné à l'amélioration de l'hévéa. *Rapp Mission IRCA, Paris, Fr.* 1974; 19 p.
17. Gonçalves PS. Expedição internacional à Amazonia no território federal de Rondônia para a coleta de material botânico de seringueira *Hevea brasiliensis*. Embrapa—CNPDS, Manaus. 1981; 59 p.
18. Lins A, Silva Gp, Nicolas D. Report of the Acre team in joint IRRDB/Brazil germplasm expedition. Embrapa/CNPDS. Embrapa/CNPDS. 1981; 24 p.
19. Paiva J. Coleta de material sexuado e assexuado nos seringais nativos do Estado de Mato Grosso. Embrapa/CNPDS, Manaus. 1981.
20. Clément-Demange A, H., Legnate H, Chapuset T, Pinard F, Seguin M. Characterisation and use of the IRRDB germplasm in Ivory Coast and French Guyana: status in 1997. Symposium on natural rubber (*Hevea brasiliensis*) Volume 1 General, soils and fertilization, and breeding and selection sessions. 1998.
21. Paiva J, Kageyama PK, Vencovsky R, Contel PB. Genetics of rubber tree [*Hevea brasiliensis* (Willd. ex Adr. de Juss.) Muell. Arg.] 1. Genetic variation in natural populations. *Silvae Genet.* 1994; 43: 306–311.
22. O. C. W, A. R. SB. The selection for high rubberwood yielding genotypes in 1995 *Hevea* germplasm, Rantau Panjang Forest Reserve. *J Biol Agric Healthc.* 2014; 4: 132–135.
23. Le Guen V, Doaré F, Weber C, Seguin M. Genetic structure of Amazonian populations of *Hevea brasiliensis* is shaped by hydrographical network and isolation by distance. *Tree Genet Genomes.* 2009; doi: [10.1007/s11295-009-0218-9](https://doi.org/10.1007/s11295-009-0218-9)
24. Schuelke M. An economic method for the fluorescent labelling of PCR fragments. *Nat Biotechnol.* 2000; 18: 233–234. PMID: [10657137](https://pubmed.ncbi.nlm.nih.gov/10657137/)
25. Peakall R, Smouse PE. GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics.* 2012; 28: 2537–2539. PMID: [22820204](https://pubmed.ncbi.nlm.nih.gov/22820204/)
26. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* 2003; 164: 1567–1587. PMID: [12930761](https://pubmed.ncbi.nlm.nih.gov/12930761/)
27. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000; 155: 945–959. PMID: [10835412](https://pubmed.ncbi.nlm.nih.gov/10835412/)
28. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour.* 2012; 4: 359–361.
29. Jakobsson M, Rosenberg NA. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* 2007; 23: 1801–1806. PMID: [17485429](https://pubmed.ncbi.nlm.nih.gov/17485429/)
30. Rosenberg NA. DISTRUCT: A program for the graphical display of population structure. *Mol Ecol Notes.* 2004; 4: 137–138.
31. Perrier X, Jacquemoud-Collet JP. DARwin software. In: <http://darwin.cirad.fr/>. 2006.
32. Schoen DJ, Brown AH. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proc Natl Acad Sci U S A.* 1993; 90: 10623–10627. PMID: [8248153](https://pubmed.ncbi.nlm.nih.gov/8248153/)
33. Lam L Van, Thanh T, Chi VTQ, Tuy LM. Genetic diversity of hevea IRRDB81 collection assessed by RAPD markers. *Mol Biotechnol.* 2009; 42: 292–298. doi: [10.1007/s12033-009-9159-7](https://doi.org/10.1007/s12033-009-9159-7) PMID: [19288228](https://pubmed.ncbi.nlm.nih.gov/19288228/)

34. Lekawipat N, Teerawatanasuk K, Rodier-Goud M, Seguin M, Vanavichit A, Toojinda T, et al. Genetic Diversity Analysis of Wild Germplasm and Cultivated Clones of *Hevea brasiliensis* Muell. Arg. by Using Microsatellite Marker. *J Rubber Res.* 2003; 6: 36–47.
35. Besse P, Seguin M, Lebrun P, Chevallier MH, Nicholas D, Lanaud C. Genetic Diversity among Wild and Cultivated Populations of *Hevea brasiliensis* Assessed by Nuclear RFLP Analysis. *Theor Appl Genet.* 1994; 88: 199–207. doi: [10.1007/BF00225898](https://doi.org/10.1007/BF00225898) PMID: [24185927](https://pubmed.ncbi.nlm.nih.gov/24185927/)
36. Yeang H, Chevallier M-H. Range of *Hevea brasiliensis* Pollen Dispersal Estimated by Esterase Isozyme. *Ann Bot.* 1999; 84: 681–684. doi: [10.1006/anbo.1999.0959](https://doi.org/10.1006/anbo.1999.0959)
37. Seguin M, Flori A, Legnate H, Clément-Demange A. Rubber Tree. In: Genetic diversity of tropical crops, Hamon P., Seguin M., Perrier X. & Glaszmann J.C JC (Ed.). *Sci Publ Enfield, USA.* 2003; 277–305.
38. Seibert RJ. The uses of *Hevea* for food in relation to its domestication. *Missouri Bot Gard Saint Louis.* 1948; 35: 61–68.
39. Baldwin JJT. *Hevea*: a first interpretation. A cytogenetic survey of a controversial genus, with a discussion of its implications to taxonomy and to rubber production. *J Hered.* 1947; 38: 54–64.
40. Saha T, Roy CB, Nazeer MA. Microsatellite variability and its use in the characterization of cultivated clones of *Hevea brasiliensis*. *Plant Breed.* 2005; 124: 86–92.
41. Souza L, Mantello C, Santos M, Souza Gonçalves P, Souza A. Microsatellites from rubber tree (*Hevea brasiliensis*) for genetic diversity analysis and cross-amplification in six *Hevea* wild species. *Conserv Genet Resour.* Springer Netherlands; 2009; 1: 75–79. doi: [10.1007/s12686-009-9018-7](https://doi.org/10.1007/s12686-009-9018-7)
42. Feng SP, Li WG, Huang HS, Wang JY, Wu YT. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). *Mol Breed.* 2009; 23: 85–97. doi: [10.1007/s11032-008-9216-0](https://doi.org/10.1007/s11032-008-9216-0)
43. Mantello CC, Suzuki FI, Souza LM, Gonçalves PS, Souza AP. Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species. *BMC Res Notes.* 2012; 5: 329. doi: [10.1186/1756-0500-5-329](https://doi.org/10.1186/1756-0500-5-329) PMID: [22731927](https://pubmed.ncbi.nlm.nih.gov/22731927/)
44. Brown AHD. Core collections: a practical approach to genetic resources management. *Genome.* 1989; 31: 818–824.
45. Richards C, Volk G. Selection of stratified core sets representing wild apple (*Malus sieversii*). *J Am Soc Hortic Sci.* 2009; 134: 228–235.
46. Balas FC, Osuna MD, Domínguez G, Perez-Gragera F, Lopez-Corralles M. *Ex situ* conservation of underutilized fruit tree species: establishment of a core collection for *Ficus carica* L. using microsatellite markers (SSRs). *Tree Genet Genomes.* 2014; 10: 703–710. doi: [10.1007/s11295-014-0715-3](https://doi.org/10.1007/s11295-014-0715-3)
47. Clément-Demange A, Legnate H, Seguin M, Carron M, Le Guen V, Chapuset T, et al. Rubber tree. In: Charrier A, Jacquot M, Hamon S, Nicolas D (eds) *Tropical plant breeding.* *Sci Publ Inc.* 2001; 455–480.

Anexo 2

Mol Breeding (2016)36:115
DOI 10.1007/s11032-016-0534-3



Development of single nucleotide polymorphism markers in the large and complex rubber tree genome using next-generation sequence data

Livia Moura de Souza · Guilherme Toledo-Silva · Claudio Benicio Cardoso-Silva · Carla Cristina da Silva · Isabela Aparecida de Araujo Andreotti · Andre Ricardo Oliveira Conson · Camila Campos Mantello · Vincent Le Guen · Anete Pereira de Souza

Received: 8 February 2016 / Accepted: 16 July 2016
© Springer Science+Business Media Dordrecht 2016

Abstract The development of single nucleotide polymorphism (SNP) markers provides the opportunity to improve many areas of plant breeding and population genetics. Unfortunately, for species such as the rubber tree (*Hevea brasiliensis*), the use of next-generation sequencing for genomic SNP discovery is very difficult because of the large genome size and the abundance of repeated sequences. Access to a set of validated SNP markers is a significant advantage for rubber researchers who wish to apply SNPs in scientific research. Here, we performed genomic sequencing of *H. brasiliensis* and generated 10,993,648 short reads, which were assembled into 10,071 contigs (N50 = 3078) by a *de novo* assembly strategy. A total of 2446 contigs presented no hits in the current *H. brasiliensis* genome assembly and may

therefore be considered novel genomic sequences of rubber tree. A total of 143 putative polymorphic positions were selected, gene annotations were available for 58.7 % of the markers, and all of the sequences could be anchored to the released *H. brasiliensis* genome. These SNPs were validated in eight genotypes of *H. brasiliensis* and 15 F1 plants from a mapping population, resulting in 30 (20.9 %) positions correctly classified. The analysis revealed key candidate genes responsible for defence mechanisms and provided markers for further genetic improvement of *Hevea* in breeding programmes.

Keywords Single nucleotide polymorphism · Next-generation sequencing · Molecular marker · *Hevea brasiliensis*

Electronic supplementary material The online version of this article (doi:10.1007/s11032-016-0534-3) contains supplementary material, which is available to authorized users.

L. M. de Souza · G. Toledo-Silva ·
C. B. Cardoso-Silva · C. C. da Silva ·
I. A. de Araujo Andreotti · A. R. O. Conson ·
C. C. Mantello · A. P. de Souza (✉)
Molecular Biology and Genetic Engineering Center
(CBMEG), University of Campinas (UNICAMP),
Campinas, SP, Brazil
e-mail: anete@unicamp.br

G. Toledo-Silva
Department of Biochemistry, Federal University of Santa
Catarina, Florianópolis, SC, Brazil

V. Le Guen
UMR AGAP, Centre de Coopération Internationale en
Recherche Agronomique pour le Développement
(CIRAD), Montpellier, Hérault, France

A. P. de Souza
Department of Plant Biology, Biology Institute,
University of Campinas (UNICAMP), Campinas, SP,
Brazil

Published online: 30 July 2016

Springer

Introduction

Until recently, single nucleotide polymorphism (SNP) marker development was expensive and time-consuming. With the advent of next-generation sequencing (NGS) and faster genotyping technologies, SNPs have emerged as the marker of choice in crop breeding (Varshney et al. 2009). SNP marker development and genotyping have provided insight into the genetics of model organisms; however, studies of non-model species have lagged behind because of the scarcity of sequences and markers. SNPs and insertions/deletions (InDels) are the most abundant types of DNA sequence polymorphisms and can be theoretically found within every genomic sequence (Rafalski 2002). SNP markers have many applications, such as cultivar identification, construction of genetic maps, assessment of genetic diversity, detection of genotype/phenotype associations, and marker-assisted breeding (Flint-Garcia et al. 2005).

There are few studies involving whole genome sequencing (WGS) efforts that have led to the successful discovery of SNPs; nonetheless, many studies in the rubber tree (*Hevea brasiliensis* (Willd. ex A.D.R. de Juss.) Muell.-Arg.) have focused on transcriptome analysis (Li et al. 2012; Mantello et al. 2014; Triwitayakorn et al. 2011). Although these RNA-seq data have added an abundance of new information on the rubber tree, the non-coding regions of the genome are also essential for understanding the regulatory elements controlling gene expression, as well as other genomic features, allowing the development of a more comprehensive set of molecular markers.

NGS technologies have also been extended to SNP discovery in large and complex genomes that lack an assembled reference genome (Bachlava et al. 2012; You et al. 2011), and additional findings from SNP research in non-model organisms are especially timely. In the case of the rubber tree, a limited number of SNPs have been described. The first study of this type on the rubber tree developed only ten SNP markers (Pootakham et al. 2011). In 2014, Salgado et al. (2014) developed 172 new SNP markers; Silva et al. (2014) characterized 43 SNP markers in 13 sequences that showed similarities to those encoding proteins involved in stress response, latex biosynthesis and developmental processes; Mantello et al. (2014) selected sequences that were identified as

belonging to the mevalonate (MVA) and 2-C-methyl-D-erythritol 4-phosphate (MEP) pathways, which are involved in rubber biosynthesis, and validated 78 SNP markers. However, there are many SNPs in non-coding regulatory regions that can be used as molecular markers, and the exact functions of these SNPs are not yet clear.

As the majority of SNPs occur in non-coding sequences of the genome, the frequency of nucleotide substitutions is almost three times higher in non-coding regions than in coding sequences (Ching et al. 2002), and their influences on phenotype may occur through biological mechanisms such as transcription factor binding and alternative splicing. Therefore, the importance of developing markers in these regions is evident for *H. brasiliensis*, which is the prime source of commercial rubber, and several other economically important species.

Hevea brasiliensis is a deciduous perennial tree of the family Euphorbiaceae. The genus *Hevea* occurs naturally only within the Amazon rainforest and its distribution extends over six million square kilometres, which is more than half of the territory of Brazil, at the boundaries of the Amazon forest. With a wide adaptation range for different ecological environments, the rubber tree exhibits considerable morphological variability, ranging from tall trees to shrubs (Gonçalves and Fontes 2012).

Due to the long generation time and the large size of the crop, new tools need to be developed to manage the germplasm bank variability and assist breeders in their recombination strategies. The development of DNA-based markers is important for the selection and improvement of varieties and hybrids in plant breeding programmes (Gupta et al. 2001; Kota et al. 2003).

We performed genomic shotgun sequencing in *H. brasiliensis* to develop SNP markers in regions of possible regulatory cis-elements of promoters and regulators. These regions are not accessible with RNA-seq; however, they are important in gene regulation and are therefore interesting targets for plant breeding programmes. This study reports the results from the characterization of rubber tree SNP markers in eight genotypes of *H. brasiliensis* and 15 F1 plants from the cross between PR255 and PB217, which were part of a mapping population.

Materials and methods

DNA preparation, library construction and sequencing

Leaves of two *H. brasiliensis* genotypes (GT1 and RRIM701 genotypes) were sampled, and their total DNA was extracted following the protocol of Doyle and Doyle (1987). DNA samples were sonicated (Bioruptor, Diogenode, Liege, Belgium) to obtain fragments ranging from 300 to 400 bp. Illumina libraries were constructed using the Paired End DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA) in 50 µl reactions containing 2 µg of DNA, as recommended by the manufacturer. The quality of the libraries was assessed on a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and samples were clustered using a TruSeq PE Cluster Kit on cBot (Illumina). Resulting libraries were sequenced on an Illumina GAIIx platform with paired-end reads of 72 bases in length using TruSeq SBS 36-Cycle kits (Illumina).

Raw data analysis and de novo assembly

The BLC files generated by Illumina sequencing from GT1 and RRIM701 genotypes were first converted to qSeq format using Off-Line Basecaller v.1.9.4 (OLB) software and then transformed into FASTQ files using a custom script. Short read data were filtered and trimmed using the CLC Genomics Workbench package (CLC Bio, Aarhus, Denmark). Reads with length <60 bases, single reads (lacking a read pair) and bases with low-quality scores ($Q < 20$) were discarded. De novo assembly was performed using CLC Genomics Workbench with the following parameters: word size (k-mer) was defined as 29, maximum gap and mismatch count were set to 2, insertion and deletion penalties were set to 3, minimum contig length was 500 bp and similarity and length fraction values were 0.8 and 0.5, respectively.

The contigs resulting from the *de novo* assembly were cleaned, removing chloroplast- and mitochondria-derived sequences. This step consisted of a similarity search against organelle sequences of *H. brasiliensis*, *Manihot esculenta*, *Ricinus communis* L., *Vitis vinifera*, *Populus trichocarpa* and *Glycine max* via BlastN using the BLAST+ suite (Camacho et al. 2009), considering a cut-off e-value of $1e-06$. The resulting dataset was used for all subsequent analysis. Selected contigs were analysed using CENSOR

(Kohany et al. 2006), a tool designed to rapidly identify repetitive elements by comparison with known repeats. The microsatellite identification tool (MISA) (Thiel et al. 2003) was used to detect simple sequence repeats (SSRs).

Variant detection

For SNP detection, reads were mapped to the filtered set of contigs to evaluate putative polymorphisms using Burrows–Wheeler Aligner (BWA) (Langmead et al. 2009). First, contigs were indexed through the index function of BWA, and then they were aligned as paired-end reads using the BWA-MEM function, using default parameters for all procedures. For variant calling, we used a strategy in which only sites detected by two different programs were selected. The mpileup pipeline of SAMtools (Li et al. 2009) was used as the first caller, with a minimum mapping quality of 30, minimum base quality of 20, minimum coverage of 10 and maximum coverage of 100. Freebayes (Garrison and Marth 2012) was used as the second caller with the following parameters: minimum read counting for variant calling 2, minimum base quality 20, minimum mapping quality 30, minimum coverage 10 and no InDels or multiple nucleotide polymorphisms (MNPs) called. Variants from sequenced genotypes were compared using the vcf-isec program from VCFtools (Danecek et al. 2011), which generated a list of unique and shared SNPs. Selection of putative polymorphic sites for validation was performed under the criteria of minimum alternate allele observations (AO) of 5, mean mapping quality of observed alternate alleles (MQM) of 50, mean mapping quality of observed reference alleles (MQMR) of 50, proportion of observed alternate alleles that are supported by properly paired read fragments (PAIRED) of 0.50 and proportion of observed reference alleles that are supported by properly paired read fragments (PAIRED) of 0.50.

Characterization of putative SNP-containing contigs

To evaluate whether novel *H. brasiliensis* genomic sequences were obtained with this *de novo* assembly, contigs were aligned to the published draft genome of *H. brasiliensis* (Rahman et al. 2013) using BlastN with a cut-off e-value of $1e-06$.

Additionally, we aimed to find contigs containing coding regions. For this purpose we used Illumina short reads of RNA-seq data from a study performed by Mantello et al. (2014). Reads from two *H. brasiliensis* genotypes (GT1 and RRIM701) were combined and then mapped to *de novo* assembled contigs using the splice-aware aligner TopHat2 (Kim et al. 2013), choosing Bowtie2 (Langmead and Salzberg 2012) as the mapper and other parameters as default. Transcripts were reconstructed from the alignment using Cufflinks 2.2.1 (Trapnell et al. 2009) with default parameters and were submitted to a BlastX search against the UniprotKB/Swiss-prot database, using BLAST+ suite (Camacho et al. 2009).

The complete coding sequences (CDS) from cassava (*Manihot esculenta*) were retrieved from <http://www.phytozome.net/cassava> (Prochnik et al. 2012) and mapped against the *de novo* assembly contigs to detect putative coding regions using sim4 software (Florea et al. 1998). BLAST hits with at least 70 % alignment were considered putative coding regions.

Manual characterization was performed for the sequences bearing the SNP markers developed here. The BLAST tool was used to search against the *H. brasiliensis* genomic sequences deposited in NCBI GenBank (Whole Genome Shotgun—WGS-database) to identify the matching sequences in the rubber tree draft genome. Annotation was performed using the BLAST tool to search against the Uniprot database (www.uniprot.org) and against the genomes of *Populus trichocarpa*, *Ricinus communis* and *Manihot esculenta* in the Phytozome database (Goodstein et al. 2012), using a cut-off e-value of $e-10$ in both databases.

KASPar assays

Only SNP positions free of other variants within the 50 bp upstream and downstream flanking sequences were selected (Table S1) for developing the KASPar assays. Furthermore, KASPar primer pairs for the targeted SNPs were ordered from KBioscience (Hertfordshire, UK); two allele-specific forward primers and one common reverse primer were designed.

Molecular markers were validated on eight genotypes of *H. brasiliensis* that are parents of studied mapping populations (GT1, PB235, RRIC100, PB217, PR255, PB260, RRIM701 and RRIM600) and 15 F1 plants from the cross between PB217 \times PR255

(Souza et al. 2013). DNA samples were extracted from lyophilized leaf tissues using a modified CTAB method (Doyle and Doyle 1987).

A 4.08 μ L KASP assay reaction contained 2 μ L KASPar 2 \times Reagent Mix (KBioscience, PN KBS-1004-001), 0.05 μ L assay mix (12 μ M each allele-specific forward primer and 30 μ M reverse primer), 0.03 μ L $MgCl_2$ (2.2 mM) and 2 μ L genomic DNA (10 ng/ μ L). The cycling conditions were as follows: 15 min at 94 $^{\circ}C$; 10 touchdown cycles of 20 s at 94 $^{\circ}C$ and 20 s at 65 $^{\circ}C$ (the annealing temperature for each cycle being reduced by 0.8 $^{\circ}C$ per cycle); and 40 cycles of 20 s at 94 $^{\circ}C$ and 60 s at 57 $^{\circ}C$. Fluorescence detection of the reactions was analysed using LightCycler 480 1.5.0 SP3 software (Roche, Basel, Switzerland).

Allele-calling data were viewed graphically as a scatter plot for each marker assayed using the LightCycler 480 II (Roche). Alleles of each SNP were scored as present, absent, duplicated or missing (failed to amplify) and converted into a binary matrix to determine minor allele frequencies (MAFs) for each SNP locus.

SNP descriptive analyses

Key descriptive statistics for measuring the informativeness of the SNP markers were calculated. The allelic polymorphism information content (PIC) and expected heterozygosity (H_e) values were calculated using the PIC calculator (<https://www.liverpool.ac.uk/~kempsj/pic.html>).

Results and discussion

Sequencing and assembling

The shotgun sequencing was a paired-end run which generated ~ 75.9 million short reads from *H. brasiliensis* DNA (GT1 and RRIM701 genotypes) using Illumina sequencing technology. The sequencing run was performed in paired-end mode, which generated reads of 72 bp (Table 1). According to Hillier et al. (2008), paired-end reads clearly increase the power to properly interpret problematic areas of the genome, including collapsed or misassembled repeats, and to detect structural variations. As genomes increase in size and complexity, paired ends

Table 1 Summary of data resulting from sequencing and the *de novo* assembly of Illumina paired-end reads from the genome of *H. brasiliensis*

<i>Sequencing</i>	
Total read count	75,820,926
Mean read length	72
Total high-quality reads	71,971,604
% high-quality reads	94.92
<i>De novo assembly</i>	
Contig count	10,071
Number of reads used	10,993,648
% reads used	14.49
Mean contig length	791
Total contig length	7,968,458
"N" occurrences in contigs	247
CpG sites	126,491
GC content in %	37.89
N50 contig length	3078
Long contigs (>10 kb)	38

will also be more efficiently placed than single-end reads. After filtering and trimming, approximately 94.92 % of data were considered high-quality (HQ) reads.

The *de novo* assembly (Table 1) used 10,993,648 short reads resulting in 10,071 contigs (Additional file 1) with a total length of 7,968,458 bp (92 % Q20 bases and 37.89 % GC content), with a mean length of 791 bp and a N50 contig length of 3078 bp. The N50 metric of *H. brasiliensis* contigs was higher than that of the sequence assembly of the barley genome (5.1 Gb), as the WGS assembly based on Illumina short reads resulted in relatively small contigs (N50 = 1425 bp) (The International Barley Genome Sequencing Consortium 2012). Direct comparisons of assembly metrics are challenging because the methods used for contig definition and/or minimum contig settings have not been standardized. Nevertheless, these metrics showed that the current assembly was successful in obtaining useful genomic contigs of *H. brasiliensis* for genomic SNP discovery.

Contig dataset filtering consisted of the removal of non-genomic-derived sequences through a homology search via BlastX against organelle sequences (chloroplast and mitochondria) of closely related species, resulting in less than 2 % (168) of contigs containing organelle sequences that were removed from the dataset. The remaining 9903 contigs were used as a

reference for subsequent analysis. First, we aimed to check whether the *de novo* assembly generated new genomic information about the species. The BlastN homology search of the assembled sequences against the *H. brasiliensis* draft genome (Rahman et al. (2013) showed that 2446 (22.7 %) of contigs were novel, while 7457 contigs exhibited high homology to existing rubber tree genomic sequences. Thus, *de novo* assembly is a useful method by which to both assemble new information and re-analyse existing public data.

Repetitive DNA was searched in the genome using CENSOR (Jurka et al. 1996), and 20.8 % of the contigs comprised repetitive sequences (Table 2). The majority of the repetitive elements were long interspersed nuclear elements and long-terminal repeat elements (LTRs, 6.4 %). Rahman et al. (2013) estimated that repeat sequences represent ~78 % of the genome, similar to the percentage in barley (84 %) (Schnable et al. 2009), and concluded that for the rubber tree less than 2 % of the total repeat elements are DNA transposons. Many repeat elements (50.24 %) could not be associated with any known family. This difference can be explained because a large percentage of reads of the repetitive regions were abandoned due to being difficult to assemble, and only 0.3 % of the estimated *Hevea* genome was sequenced in this work, which is likely not sufficient to represent the complexity of the genome.

MISA analysis identified 4684 putative microsatellites (Table 2). The most abundant repeat motifs were mononucleotides (1674), followed by dinucleotides (1157) and tetranucleotides (777). NGS clearly offers a rapid means of acquiring the sequences needed to detect SSRs.

SNP calling and molecular marker validation

To detect putative variants in the dataset, we mapped all reads from the GT1 and RRIM701 sequenced genotypes onto the *de novo* assembled contigs (10,071 contigs) using Burrows–Wheeler Aligner (BWA). Approximately 8.2 million reads were successfully mapped, resulting in ~35× depth (Table 3). Using the mapping file as a reference, two different strategies (SAMtools mpileup and Freebayes) were applied to uncover putative SNP variants (Table 4).

A total of 57,820 and 6221 putative SNPs were called by the Freebayes and SAMtools methodologies,

Table 2 Main classes of repeat elements identified in the *H. brasiliensis* genome assembly

Repeat class	Fragments	Total length (bp)
Integrated virus	2145	16,498
DNA transposon	1578	209,909
LTR retrotransposon	2314	509,096
Non-LTR retrotransposon	385	59,409
Transposable element	4297	782,094
Repetitive element		
Simple sequence repeat (SSR)	4684	83,504
Total	15,403	1,660,510

Table 3 Summary of BWA mapping of high-quality reads onto the *de novo* assembled *H. brasiliensis* contigs

Metrics	Results
Total reads	71,971,604
Mapped reads	8,223,980
% of mapped reads	11.42
Both pairs mapped	4,474,291
Singletons	3,749,689
Average depth	35.40X

Table 4 Summary of SNP calling in *de novo* assembled *H. brasiliensis* contigs

Method/metric	No. of SNPs
Freebayes (1)	57,820
SAMtools (2)	6221
Intersection of (1) and (2)	3779
Transitions (Ts)	2546
Transversions (Tv)	1233
Ts/Tv	2.06
SNP frequency	1 SNP per 2.1 kb

respectively. In search of a more conservative approach, we used the intersection between the two strategies, resulting in a final SNP dataset of 3779 variants, corresponding to an average occurrence of one SNP for every 2.1 kb. Pootakham et al. (2011) and Salgado et al. (2014) detected one SNP per 1.5 kb and one SNP per 5.2 kb for the rubber tree, respectively, using Roche 454 pyrosequencing technology, whereas Mantello et al. (2014) used Illumina sequencing technology and detected an average of one SNP per 125 bp. It is difficult to make a comparison because the methodologies are different with regard to

sequencing technique, sequenced material (genomic DNA vs. mRNA) and detection of SNPs.

A total of 2546 transitions and 1233 transversions were detected, with A↔G being the most common variation (1304; 34.48 %) and A↔C the least common variation (263; 6.9 %) observed (Table 4). The ratio of the number of transitions (Ts) to the number of transversions (Tv) is particularly helpful for assessing the quality of SNP calls (DePristo et al. 2011). The Ts/Tv ratio in our study was 2.06, indicating that the conservative approach resulted in higher accuracy regarding variant detection.

Considering the total of 3779 SNPs that were filtered, only those free of other variants within the 50 bp upstream and downstream regions flanking the central base position were selected. Based on this strategy, a total of 143 putative polymorphic positions were selected for further evaluation. Putative SNPs were validated on eight genotypes of *H. brasiliensis* which are parents of currently studied mapping populations (GT1, PB235, RRIC100, PB217, PR255, PB260, RRIM701 and RRIM600) and 15 F1 plants from the cross between PB217 × PR255 (Souza et al. 2013). Thirty (20.9 %) positions presented profiles that could be correctly classified and worked well as potential molecular markers (Additional file 2 Table S1). An example of a correctly classified SNP is the segregation of the locus sHbUNI497 (Fig. 1a), where PB217 was classified as heterozygous (XY), PR255 was classified as homozygous (XX), and the F1 genotypes segregation presented two classes (XX and XY) as expected.

Of the positions, 31 were monomorphic across the samples and 48 were no-call SNPs (failed to amplify). These monomorphic markers may have resulted from errors in sequencing, which then led to the misidentification of the SNP, and some genotyped SNPs may

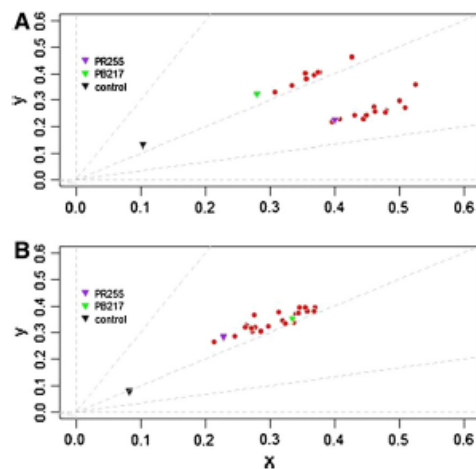


Fig. 1 SNP assays using the Fluidigm platform in parents of mapping populations and 15 hybrids (red colour). **a** Expected segregation of sHbUNI497 and **b** likely duplication segregation of sHbUNI294

have failed due to the presence of occult variants (Bentley et al. 2008) or due to poor PCR amplification and low signal intensities resulting in missing data.

The mean PIC was 0.283, ranging from 0.111 to 0.368. The mean expected heterozygosity was 0.353, ranging from 0.118 to 0.486 (Additional file 2 Table S1). These values are lower than those found by Mantello et al. (2014). However, it should be noted that the present set of markers was analysed on a limited number of genotypes (parents of mapping populations), while Mantello et al. (2014) screened 36 genotypes from a germplasm bank of *Hevea*.

Approximately 23.77 % (34) of the 143 markers analysed showed a segregation profile in which all analysed genotypes were heterozygotes, and were considered a duplication (Additional file 2 Table S1). All genotypes were heterozygotes (including the F1 genotypes from PB217 × PR255 crossing) at the locus sHbUNI294, which can be observed in Fig. 1b. The same results were obtained by Salgado et al. (2014); ten SNPs displaying the same heterozygous combination for all the genotypes were found. This scenario was not expected, considering the parental origin and because some hybrids of the mapping populations were analysed. Our hypothesis is that this segregation profile was caused by the presence of duplicated regions throughout the *H. brasiliensis* genome. These

duplicated regions may have gone through independent mutations over the course of evolution, although they maintain homozygosity in each duplicated locus.

The corresponding reads of these regions were mapped as if they were from the same locus, showing a SNP-type variation in the in silico analysis. In the KASPar technique, both loci would have been amplified; thus, each locus would be responsible for 50 % of the amplicons, which created a heterozygous profile of all genotypes evaluated.

Locus duplication, as revealed by microsatellite molecular markers, has been reported in *H. brasiliensis* and other *Hevea* sp. (Le Guen et al. 2011; Mantello et al. 2012; Silva et al. 2014; Souza et al. 2009); however, such evidence highlights the importance of focusing efforts on elucidating these possible duplications.

Functional annotation and SNP marker characterization

To improve the quality of annotations, we aligned ~94 million paired-end reads from RNA-seq data (Mantello et al. 2014) to the 10,071 contigs using TopHat2 (Kim et al. 2013), which allows gaps in the read-to-reference alignment at putative splice sites. Approximately 3.8 % of the total reads were successfully mapped on the *de novo* assembled genomic regions. Cufflinks was used to aid the refinement of gene structures by creating transcript “fragments” with sharply defined exon boundaries (Trapnell et al. 2009). Cufflinks reconstruction yielded 4374 different transcripts belonging to 3934 assembled genomic contigs.

The total set of contigs was aligned against the cassava (*Manihot esculenta*) complete CDS using sim4 software (Florea et al. 1998). We identified 1447 contigs with predicted coding regions, with at least 70 % of the correspondence in the alignment (Additional file 2 Table S2).

A large number of molecular markers isolated from genomic DNA during the last few decades have been located in the intergenic or genic regions of the genome without any information available on their functions (Varshney et al. 2007). To evaluate the utility of the 143 SNP markers developed, the contig sequences of these markers were manually annotated and compared using the BLAST tool to search against the Uniprot database (www.uniprot.org) and against

the genomes of *Populus trichocarpa*, *Ricinus communis* and *Manihot esculenta* in the Phytozome database (Additional file 2 Table S3).

All contigs containing SNP markers showed significant sequence homology to the draft of the *Hevea* genome (Rahman et al. 2013) (Additional file 2 Table S4). From the 143 SNP markers developed, 86 SNP markers (60.1 %) are inserted in sequences that presented similarities to those encoding proteins with known functions. The majority of these sequences are involved in metabolic processes (27 sequences—e.g.: sHbUNI409), followed by developmental processes (19—e.g.: sHbUNI289, sHbUNI399) and stress response (19—e.g.: sHbUNI371, sHbUNI419). Other identified functions were transport (6—e.g.: sHbUNI380), cell cycle (4—e.g.: sHbUNI495), gene silencing (3—e.g.: sHbUNI377) and respiration (2—sHbUNI403).

sHbUNI402 was identified in contig_4931, which was annotated in Swiss-prot as a probable disease-resistance protein (*Arabidopsis thaliana* At4g27220) (Additional file 2 Table S3). In contig_5088, the SNP marker sHbUNI419 was associated with a disease-resistance protein (TMV resistance protein in *Nicotiana glutinosa*).

For the sHbUNI439 marker sequence, there was no matching protein in either the Uniprot or Phytozome databases; nevertheless, the sequence was highly similar to genomic sequences from two *M. esculenta* scaffolds (scaffold05875 and scaffold12525, e-values = 0 and $3.4e-101$, respectively) and one scaffold from the *R. communis* genome (29706, e-value = $8e-46$). In all three scaffolds the matching sequences were close to transcripts that are similar to telomeric repeat binding protein (TRBP) 1. In *R. communis*, this protein is encoded by a single copy gene, while in *M. esculenta* there are three copies (Phytozome database, as of December 2015). Moreover, *M. esculenta* scaffold12525 presents a structure similar to *R. communis* scaffold29706 around the matching sequence; however, scaffold05875 of *M. esculenta* differs from the previous scaffolds. Because it is a genomic region that is conserved among different species from the same family and is close to transcripts that are similar to the same protein, it is possible that the region in which marker sHbUNI439 is inserted is a cis-element that regulates the expression of a similar protein in *H. brasiliensis*.

Of the SNP markers developed, 34 exhibited a profile which suggested that the loci analysed were duplicated. The sequences being analysed may be members of conserved gene families, or the regions may in fact have been duplicated. For sHbUNI320, the sequence used for the assay design of this marker presented similarity to an Argonaute (AGO) protein. AGO proteins are essential for small RNA silencing pathways, as they bind to different siRNAs and miRNAs and mediate the repression of specific RNAs by degradation or translation inhibition (Höck and Meister 2008). The AGO protein family has several members in different plant species, including *Arabidopsis thaliana* (10), *Populus trichocarpa* (15), *Ricinus communis* (9) and *Manihot esculenta* (13) (Phytozome database, as of December 2015). A possible explanation for the duplicated profile that marker sHbUNI320 presented is that two members of the AGO family were amplified by the same primers.

Likewise, marker sHbUNI324 presented a duplicated profile; however, the sequence in which this SNP was detected presented similarity to that encoding chromosome transmission fidelity 7 (Ctf7) protein, which is a single gene in the majority of plant species (Phytozome database, as of December 2015). Ctf7 is required for the establishment of sister chromatid cohesion, regulation of chromosome segregation and DNA repair. Although Ctf7 is a single-copy gene in most plant genomes, the presence of another copy of this sequence in the rubber tree genome is possible. Another possibility is that the designed primers for the sHbUNI324 marker amplified another HSF1 TFBS region in the *H. brasiliensis* genome.

This study demonstrates that high-throughput DNA sequencing is a powerful approach for the identification of novel sequences and the rapid development of SNP markers in non-model organisms. In addition, new SNP markers possibly involved with different plant molecular mechanisms, such as plant responses to abiotic and biotic stresses and plant developmental processes, were developed. These SNPs are an extremely useful source of markers in rubber tree breeding for marker-assisted selection and gene-based cloning.

The new SNP markers identified in this work provide an excellent tool for the enrichment of genetic linkage maps and the identification of candidate genes for traits of interest (QTL) in progeny from crosses

developed by breeding programmes, as well as for studying genetic diversity in the rubber tree.

Acknowledgments The authors gratefully acknowledge the Fundação de Amparo a Pesquisa do Estado de São Paulo, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Computational Biology Program and Agropolis Program) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico for financial support and scholarships and a research fellowship.

Authors' contributions LMS performed the molecular genetic studies, helped to perform the biocomputational analysis and drafted the manuscript. GTS, CBCS and CCS performed a biocomputational analysis and drafted the manuscript. GTS, ARC, CCM and IAAA assisted in the molecular genetics studies. VLG participated in the evaluations of the molecular data and helped to draft the manuscript. APS conceived the study, participated in its design and coordination and helped to draft the manuscript. All of the authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Bachlwa E, Taylor CA, Tang S, Bowers JE, Mandel JR, Burke JM, Knapp SJ (2012) SNP discovery and development of a high-density genotyping array for sunflower. *PLoS One* 7:e29814. doi:10.1371/journal.pone.0029814
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59. doi:10.1038/nature07517
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinform* 10:421. doi:10.1186/1471-2105-10-421
- Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet* 3:19–32
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158. doi:10.1093/bioinformatics/btr330
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498. doi:10.1038/ng.806
- Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11–15
- Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* 44:1054–1064. doi:10.1111/j.1365-3113.2005.02591.x
- Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 8:967–974
- Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. Preprint at [arXiv:1207.3907](https://arxiv.org/abs/1207.3907)[q-bio.GN]
- Gonçalves PS, Fontes JRA (2012) Domestication and breeding of the rubber tree. In: Borém A, Lopes MTG, Clement CR, Noda H (eds) Domestication and breeding: Amazon species. UFV, Viçosa, pp 393–420
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186. doi:10.1093/nar/gkr944
- Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr Sci* 80:524–535
- Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JL, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 5:183–188. doi:10.1038/nmeth.1179
- Höck J, Meister G (2008) The Argonaute protein family. *Genome Biol* 9:210. doi:10.1186/gb-2008-9-2-210
- Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem* 20:119–121. doi:10.1016/S0097-8485(96)80013-1
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14:R36. doi:10.1186/gb-2013-14-4-r36
- Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: repbaseSubmitter and censor. *BMC Bioinform* 7:474. doi:10.1186/1471-2105-7-474
- Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Mol Genet Genomics* 270:24–33. doi:10.1007/s00438-003-0891-6
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nat Methods* 9:357–359. doi:10.1038/nmeth.1923

- Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) Searching for SNPs with cloud computing. *Genome Biol* 10:R134. doi:10.1186/gb-2009-10-11-r134
- Le Guen V, Gay C, Xiong TC, Souza LM, Rodier-Goud M, Seguin M (2011) Development and characterization of 296 new polymorphic microsatellite markers for rubber tree (*Hevea brasiliensis*). *Plant Breed* 130:294–296. doi:10.1111/j.1439-0523.2010.01774.x
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. doi:10.1093/bioinformatics/btp352
- Li D, Deng Z, Qin B, Liu X, Men Z (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics* 13:192. doi:10.1186/1471-2164-13-192
- Mantello CC, Suzuki FI, Souza LM, Gonçalves PS, Souza AP (2012) Microsatellite marker development for the rubber tree (*hevea brasiliensis*): characterization and cross-amplification in wild *hevea* species. *BMC Res Notes* 5:329. doi:10.1186/1756-0500-5-329
- Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, de Souza Gonçalves P, Vicentini R, de Souza AP (2014) De novo assembly and transcriptome analysis of the rubber tree (*hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. *PLoS One* 9:e102665. doi:10.1371/journal.pone.0102665
- Pootakham W, Chanprasert J, Jomchai N, Sangsrukru D, Yoocha T, Therawattanasuk K, Tangphatsomruang S (2011) Single nucleotide polymorphism marker development in the rubber tree, *hevea brasiliensis* (Euphorbiaceae). *Am J Bot* 98:e337–e338. doi:10.3732/ajb.1100228
- Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S (2012) The cassava genome: current progress, future directions. *Trop Plant Biol* 5:88–94
- Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* 5:94–100. doi:10.1016/S1369-5266(02)00240-6
- Rahman AY, Usharaj AO, Misra BB, Thottathil GP, Jayasekaran K, Feng Y, Hou S, Ong SY, Ng FL, Lee LS, Tan HS, Sakaff MK, Teh BS, Khoo BF, Badai SS, Aziz NA, Yuryev A, Knudsen B, Dionne-Laporte A, Mchunu NP (2013) Draft genome sequence of the rubber tree *hevea brasiliensis*. *BMC Genomics* 14:75. doi:10.1186/1471-2164-14-75
- Salgado LR, Koop DM, Pinheiro DG, Rivallan R, Le Guen V, Nicolás MF, de Almeida LG, Rocha VR, Magalhães M, Gerber AL, Figueira A, Cascardo JC, de Vasconcelos AR, Silva WA, Coutinho LL, Garcia D (2014) De novo transcriptome analysis of *Hevea brasiliensis* tissues by RNA-seq and screening for molecular markers. *BMC Genomics* 15:236. doi:10.1186/1471-2164-15-236
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115. doi:10.1126/science.1178534
- Silva CC, Mantello CC, Campos T, Souza LM, Gonçalves PS, Souza AP (2014) Leaf-, panel- and latex-expressed sequenced tags from the rubber tree (*hevea brasiliensis*) under cold-stressed and suboptimal growing conditions: the development of gene-targeted functional markers for stress response. *Mol Breed* 34:1035–1053. doi:10.1007/s11032-014-0095-2
- Souza LM, Mantello CC, Santos MO, de Souza Gonçalves P, Souza AP (2009) Microsatellites from rubber tree (*Hevea brasiliensis*) for genetic diversity analysis and cross-amplification in six *hevea* wild species. *Conserv Genet Resour* 1:75–79. doi:10.1007/s12686-009-9018-7
- Souza LM, Gazaffi R, Mantello CC, Silva CC, Garcia D et al (2013) QTL mapping of growth-related traits in a full-sib family of rubber tree (*Hevea brasiliensis*) evaluated in a sub-tropical climate. *PLoS One* 8:e61238. doi:10.1371/journal.pone.0061238
- Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 106:411–422. doi:10.1007/s00122-002-1031-0
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111. doi:10.1093/bioinformatics/btp120
- Triwitayakorn K, Chatkulkawin P, Kanjanawattana Wong S, Sraphet S, Yoocha T, Sangsrukru D, Chanprasert J, Ngamphiw C, Jomchai N, Therawattanasuk K, Tangphatsomruang S (2011) Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. *DNA Res* 18:471–482. doi:10.1093/dnares/dsr034
- Varshney RK, Beier U, Khlestkina EK, Kota R, Korzun V, Graner A, Börner A (2007) Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency, and applications for genome mapping and diversity studies. *Theor Appl Genet* 114:1105–1116. doi:10.1007/s00122-007-0504-6
- Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 27:522–530. doi:10.1016/j.tibtech.2009.05.006
- You FM, Huo N, Deal KR, Gu YQ, Luo M-C, McGuire PE, Dvorak J, Anderson OD (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12:59. doi:10.1186/1471-2164-12-59

Anexo 3

Profa. Dra. Rachel Meneguello
Presidente
Comissão Central de Pós-Graduação
Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Mapeamento Genético-Molecular e Análise de QTLs associados ao crescimento em Hevea brasiliensis utilizando marcadores microssatélites e SNPs**, não infringem os dispositivos da Lei n.º 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 20 de outubro de 2016

Assinatura : André R. O. Conson
Nome do(a) autor(a): **André Ricardo Oliveira Conson**
RG n.º 94169151

Assinatura : Anete Pereira de Souza
Nome do(a) orientador(a): **Prof.^a Dr.^a Anete Pereira de Souza**
RG n.º 8680325

Anexo 4



COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br



DECLARAÇÃO

Em observância ao §5º do Artigo 1º da Informação CCPG-UNICAMP/001/15, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "***Mapeamento Genético-Molecular e Análise de QTLs associados ao crescimento em Hevea brasiliensis utilizando marcadores microssatélites e SNPs***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: André R. O. Conson
Nome do(a) aluno(a): André Ricardo Oliveira Conson

Assinatura: Anete Pereira de Souza
Nome do(a) orientador(a): Prof.^a Dr.^a Anete Pereira de Souza

Data: 20/10/2016